Estimation and Hypothesis Testing for Stochastic Differential Equations with Time-Dependent Parameters

by

Yanqiao Zhang

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2012

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Yanqiao Zhang

Abstract

There are two sources of information available in empirical research in finance: one corresponding to historical data and the other to prices currently observed in the markets. When proposing a model, it is desirable to use information from both sources. However in modern finance, where stochastic differential equations have been one of the main modeling tools, the common models are typically different for historical data and for current market data. The former are usually assumed to be time homogeneous, while the latter are typically time in-homogeneous. This practice can be explained by the fact that a time-homogeneous model is stationary and easier to estimate, while time-inhomogeneous model are required in order to replicate market data sufficiently well without creating arbitrage opportunities.

In this thesis, we study methods of statistical inference, both parametric and non-parametric, for stochastic differential equations with time-dependent parameters. In the first part, we propose a new class of stochastic differential equation with time-dependent drift and diffusion terms, where some of the parameters change according to a hidden Markov process. We show that under some technical conditions this innovative way of modeling switching times renders the resulting model stationary. We also explore different approaches to estimate parameters in our proposed model. Our simulation studies demonstrate that the parameters of the model can be efficiently estimated by using a version of the filtering method proposed in the literature. We illustrate our model and the proposed estimation method by applying them to interest rate data, and we detect significant time variations in early 1980s, when targets of the monetary policy in the United States were changed.

One of the known drawbacks of parametric models is the risk of model misspecification. In the second part of the thesis, we allow the drift to be time-

dependent and nonparametric, and our objective is to estimate it using a single trajectory of the process. The main idea underlying this method is to approximate the time-dependent function with a sequence of polynomials. Since we can estimate efficiently only a finite number of parameters for any finite length of data, in our method we propose to relate the number of parameters to the length of the observed trajectory. This idea is similar to the method of sieves proposed by Grenander (Abstract Inference, 1981). The asymptotic analysis that we present is based on the assumption that the length of available data T increases to infinity. We investigate two cases, one is a Brownian motion with time-dependent drift and the other corresponds to a class of mean-reverting stochastic differential equations with time-dependent mean-reversion level. In both cases we prove asymptotic consistency and normality of a modified maximum likelihood estimator of the projected time-dependent component. The main challenge in proving our results in the second case stems from two features of the problem: one is due to the fact that coefficients of projections change with T and the other is related to the confounding effect between the mean-reversion speed and the level function. By applying our method to the same interest rate data we use in the first part, we find another evidence of time-variation in the drift term.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Adam Kolkiewicz, for his kindness, support, patience and encouragement. Without his help, the thesis would not have been possible. I would also like to show my gratitude to my thesis committee members: Professors John Knight, Don McLeish, Ken Vetzal and Tony Wirjanto, for carefully reading my thesis and providing insightful comments and suggestions.

I would like to thank Professor Chengguo Weng, Zhaoxia Ren and Zhongxian Men for rehearsing my presentation and providing helpful advices. I would also like to thank Mary McPherson, who has kindly proofread my entire thesis during a relatively short time. Special thanks are given to Wenjing Ge for always being there to help me out, saving me time to focus on my thesis, and hence making my life easier when I felt frustrated and stressed. My thanks are also given to Dr. Houben Huang and Dr. Anthony Vaz, who encouraged me to finish the degree while my working at Bank of Montreal and Manulife Financial, respectively.

I am indebted to many of my friends and colleagues who helped me in various ways during the course of my PhD study. The following list is not meant to be exhaustive: Yonathon Beraki, Zhenyu Cui, Yasaman HosseinKashi, Yue Jin, Chao Qiu, Reza Ramezan, Ying Shang, Hua Shen, Basil Singer, Kai Wang, Wei Wei, Longyang Wu, Yuxin Zhang, Hui Zhao and Ming Zhou.

Last, but not the least, I would like to dedicate this thesis to my parents, sister and brother, who have always been there whenever I needed support.

Contents

List of Tables

List of Figures

Chapter 1

Background and Motivation

1.1 Introduction to the Problem

Stochastic processes are widely used for model building in the social, physical, engineering, and life sciences as well as in financial economics. Some well-known studies on foundational probabilistic knowledge in stochastic processes include those of Lipster and Shiryaev (2001, 2010), Karatzas and Shreve (1991), Kloeden and Platen (1992). Statistical inference for stochastic processes is of great importance from the theoretical as well as from applications point of view in model building. During the past three decades, statistical inference for stochastic processes has been extensively studied. Bishwal (2008), for instance, presents the estimation of unknown parameters in corresponding continuous models based on continuous and discrete observations and extensively examines maximum likelihood, minimum contrast, and Bayesian methods. Kutoyants (2004) focuses on inference theory for ergodic diffusion processes. Prakasa Rao (1999) brings together several methods of estimating the parameters involved in diffusion-type processes for data sets that are discretely or continuously sampled. The work in my thesis is motivated by the

applications of stochastic processes in finance and focuses primarily on statistical inference.

In finance, two sources of information are typically available: historical data, which reflects the property of underlying securities under an objective measure (the P measure), and market data, such as prices of options and futures on the same security under a risk-neutral measure (the Q measure). In other words, the P measure is the probability measure implied by realized historical data of the underlying security itself, while the Q measure is the one chosen by current market to price derivatives of the underlying security. Therefore, by nature, the P measure contains information about the past, and the Q measure contains information about the future, as derivative prices can be calculated as discounted expected value of future payoff under this measure. Using information from both sources is advisable for identifying a model of the process; indeed, some work has been done in volatility modeling of asset returns to combine information from both P and Q measures. For example, Chernov (2001) argues that the volatility risk premium, which accounts for the difference between implied volatility (under the Q measure) and realized volatility (under the P measure), can be used to explain why implied volatility can be a biased estimator of future realized volatility. On the other hand, Gospodinov, Gavala and Jiang (2006) find that implied volatility serves as an unbiased estimate of future volatility. The authors compare the prediction powers of different models, including conditional mean models that model implied volatility using information from the Q measure and conditional volatility models that make use of information from the P measure. The conditional volatility models include EGARCH, FIEGARCH, and stochastic volatility models. They propose an intercept correction method which can significantly improve the forecast of average integrated volatility for conditional volatility models.

To the best of our knowledge, a reconciliation of models estimated under the P measure and those estimated under the Q measure has not been done in interest

rate modeling (short rate models). This situation very likely arises because the parametrization used for estimation from historical data and that used for calibration to market data are very different. The parametrization used for fitting historical data is usually in the form of time-homogeneous models, such as those of Vasicek (1977); Cox, Ingersoll, and Ross (CIR) (1985); Chan, Karolyi, Longstaff, and Sanders (CKLS) (1992) and Stanton (1997). Such models are typically defined as solutions to stochastic differential equations. For example,

$$\begin{aligned}
\text{Vasicek (1977): } dr(t) &= k(\theta - r(t))dt + \sigma dW(t) && (1.1)\\
\text{CIR(1985): } dr(t) &= k(\theta - r(t))dt + \sigma\sqrt{r(t)}dW(t) && (1.2)\\
\text{CKLS (1992): } dr(t) &= (\alpha_0 + \alpha_1 r(t))dt + \sigma r(t)^{\gamma} dW(t), && (1.3)
\end{aligned}$$

where $\{W(t)\}$ is a standard Brownian motion. Later we refer to these models as Vasicek, CIR and CKLS models or processes. Time-homogeneous models are used mainly for understanding the behavior of interest rate and making forecasts. One main drawback of time-homogeneous models is that they are not able to fit exactly the currently observed term structure of interest rate, which is a crucial component for pricing and hedging interest rate derivatives. Therefore, the parameterization of models calibrated to market prices is often time dependent. For example, Ho and Lee (1986); Hull and White (1994a); Black et al. (1990); and Black and Karasinski (1991) have proposed the following models for pricing interest rate options:

$$\begin{aligned}
\text{HL: } dr(t) &= \mu(t)dt + \sigma(t)dW(t) && (1.4)\\
\text{HW: } dr(t) &= [\nu(t) - ar(t)]dt + \sigma dW(t) && (1.5)\\
\text{BDT: } dr(t) &= \{\alpha_1(t)r(t) + \alpha_2(t)r(t)\log(r(t))\}dt + \beta_0(t)dW(t) && (1.6)\\
\text{BK: } dr(t) &= \{\alpha_1(t)r(t) + \alpha_2(t)r(t)\log(r(t))\}dt + \beta_0(t)r(t)dW(t), && (1.7)
\end{aligned}$$

where the functions $\mu(t), \sigma(t), \nu(t), \alpha_1(t), \alpha_2(t), \beta_0(t)$ must be determined in the calibration process (the exact form of $\nu(t)$ is given in (1.12)). Time-dependent models used for pricing are generally not stationary, which makes estimating components of such models from historical data quite difficult.

Below, we review some existing works on statistical modeling for time-dependent processes. For economic time series, many papers address time-varying features of the mean. For example, Ashley and Patterson (2010) show that smooth variation in the mean induces apparent long memory; Cogley and Sbordone (2008) study time-dependent trends in the inflation rate and find that inflation persistence results mainly from the variation in the long-run trend component of inflation. The proposed estimators of the time-varying mean include a nonparametric nonlinear trend regression as in Ashley and Patterson (2010), the simple moving average as in the "moving mean" model in Ashley and Patterson (2007), and a sophisticated nonlinear bandpass filter as in Baxter and King (1999).

In the context of continuous stochastic processes, Dehling et al. (2010) study the following SDE:

$$dX(t) = (L(t) - \alpha X(t))dt + \sigma dW(t).$$

The authors assume that the function $L(t) = \sum_{i=1}^{n} \mu_i \phi_i(t)$ is a periodic and parametric function, and prove asymptotic consistency and normality of $\hat{\mu}_i$'s as $T \to \infty$. Beder(1987) presents a sieve estimator of the mean function $m(t)$ of a general Gaussian process. The author proves that his estimator is asymptotically unbiased and consistent at each t when the number of independent trajectories $n \to \infty$. This asymptotic scheme for sieve estimators has also been studied in Prakasa Rao (2004), Stone and Huang (2003), Nguyen and Pham (1982), and Geman and Hwang (1982), etc. Wirjanto (2010) discusses a nonparametric test statistic for the diffusion coefficient of the following SDE,

$$dX(t) = \mu(X(t))dt + \sigma(X(t))dW(t).$$

The test statistic is derived using an empirical-likelihood method and a nonparametric estimator of the diffusion coefficient that is asymptotically independent of the drift coefficient. Using a certain transformation of the diffusion process, the

author also discusses how the nonparametric test can be extended to the following time-inhomogeneous SDE,

$$dX(t) = \mu(X(t), t)dt + \sigma(X(t), t)dW(t).$$

Below we review in greater detail two papers that study time-dependent diffusion processes in the finance area. They were the main motivation behind our devoting this thesis to statistical inference for time-dependent SDEs.

In Fan et al. (2003), a family of time-dependent diffusion processes has been proposed to model interest rate. The authors have introduced nonparametric methods, based on local constant fitting, to estimate time-varying effects in the drift and diffusion coefficients in the following model:

$$dX(t) = [\alpha_0(t) + \alpha_1(t)X(t)]dt + \beta_0(t)X(t)^{\beta_1(t)}dW(t), \tag{1.8}$$

which encompasses most of the popular continuous diffusion models in the literature, such as in equations (1.1), (1.2), and (1.3). The authors have applied their methods to weekly US treasury bill data and found no evidence of time-inhomogeneity in the drift coefficient. However, the accuracy of their estimates suffers from the low efficiency of nonparametric estimation methods.

Al-Zoubi (2009) has studied short-term interest rate using monthly three-month T Bill data from January 1934 to July 2002. The short rate model is prescribed as a nonlinear trend stationary process, which is a sum of a time-dependent function and a CKLS process:

$$r(t) = \eta(t) + \epsilon(t) \tag{1.9}$$

$$d\epsilon(t) = [\alpha + \beta\epsilon(t)]dt + \sigma\epsilon^{\gamma}(t)dW(t), \tag{1.10}$$

where $\eta(t)$ is specified as a Chebyshev polynomial (Appendix 4.7.1). After filtering the nonlinear signal $\eta(t)$ using the ordinary least squares method, the author fitted

the stationary component with a CKLS model and estimated parameters α, β, σ and γ using the GMM method proposed in Hansen (1982). The author finds that the goodness-of-fit improves significantly for those models with drift-induced mean reversion and worsens for those with high volatility elasticity. However, the author's methodology has not been fully justified in the paper. Below we list some issues:

1) The choice of the order m of the polynomial function $\eta(t)$ is crucial. When the order of the fitted polynomial is too high, the observed trajectory will be over-fitted. As a result, the residuals will fluctuate around zero and appear to be mean-reverting. To illustrate the over-fitting problem, we give one example. Figure 1.1 shows simulated daily observations of a Brownian motion with diffusion coefficient 2 and residuals after fitting a polynomial of order 10. It can be seen that the residuals fluctuate around zero and look like a mean-reverting process. Therefore, a misspecified value of m could lead to misleading conclusions. An objective procedure to determine the order of the polynomial is desirable. In the paper, the author chooses $m = 10$, but no rationale is given for this selection. Accordingly, the conclusion in the paper may need to be drawn with further caution. The fact that this polynomial order selection is an important issue can be supported by the following sentence (Page 53, Bierens, 1997): "A more difficult problem is to provide general guidelines for specifying the order m of the detrended Chebyshev polynomials [...] Another approach is to let m converge to infinity with the sample size n at some controlled rate, [...] Whether such a sequence m_n exists, and if so how it depends on n, is an open question."

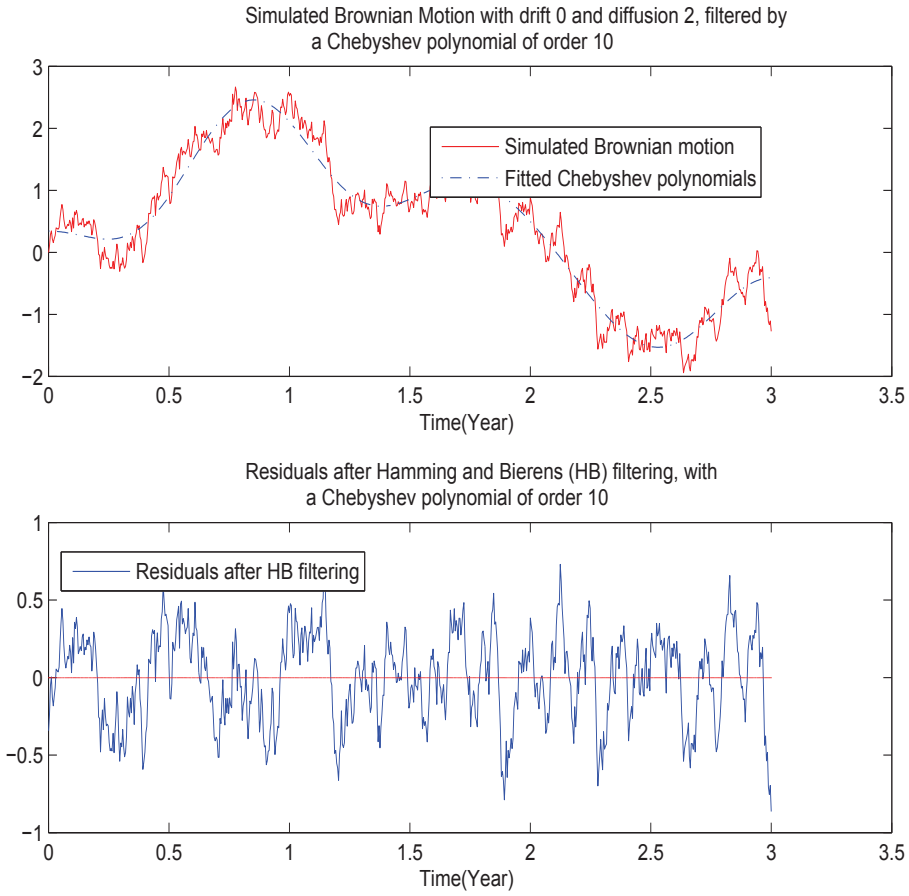


Figure 1.1: Simulated Brownian motion with drift 0 and diffusion 2, filtered by a Chebyshev polynomial of order 10.

2) Suppose for now that the polynomial order m is correctly selected. The author filters the polynomial trend using the ordinary least squares method and then estimates the parameters α, β, σ and γ in model (1.10) from the residuals. In our opinion, this two-step estimation approach will introduce more estimation error, since the process $\{\epsilon(t)\}$ is not observable and hence the residuals after filtering are not a genuine CKLS process. The main idea is that the ordinary least square method does not take into account the true nature of the residual process, such as the correlation structure of the residuals. This observation

will further increase the estimation error presented in the paper and cast doubts on the conclusions.

3) The following line is found in the paper:"After filtering the interest-rate data and rescaling, we use the GMM estimation ... to capture the stationary component of the interest rate." A CKLS model is usually applied to a positive data process, while least squares fitting of any data always results in residuals fluctuating around zero. Therefore, the author "rescales" the residuals before applying the CKLS model. However, no details are given on how the "rescaling" method is conducted and what the justification is behind this approach.

This thesis is devoted to statistical inference for time-dependent stochastic differential equations (SDEs) under P measure and consists of two parts: in the first half of the thesis (Chapters 2 and 3), we propose a new class of time-dependent SDEs by incorporating a latent continuous-time Markov chain and allowing the observed process to be dependent on the elapsed time after the state of the Markov chain is switched. Our results show that such models can be efficiently estimated. However, the time-dependent component of these models must be correctly specified.

In the second half of thesis (Chapters 4 and 5), we address the issue of model specification of the time-dependent component by proposing a method for estimating the nonlinearity in the drift. We assume that only one realization of the process is available and point out limitations of inference methods resulting from this assumption. In the proposed method, we identify the time-dependent component of the drift that can be estimated at a given level of accuracy.

Below we describe in greater detail each part and our contribution.

Part I: Parametric inference for time-dependent SDEs

Stationarity is a common assumption in stochastic modeling and states that finite dimensional distributions of the process do not change over time. Therefore, more observations of the process imply more information about the stationary distribution of the process, and hence our estimation of parameters in the model becomes more accurate. In the context of estimation, the stationary stochastic processes are typically assumed to be time homogeneous (see, for example, Kutoyants (1984) and Stanton (1997)). On the other hand, time-homogeneous models have limitations because they are not capable of capturing the time effect. These limitations have been documented, for example, by Fan et al. (2003) and Cai and Hong (2009). In the context of pricing, a variety of time-dependent diffusion models have been proposed, as described earlier in this section. Such models are typically non-stationary, due to their time-varying parameters, and therefore may pose challenges in statistical inference methods. For example, for the model (1.8) proposed in Fan et al. (2003), the asymptotic properties of the resulting estimators are still unknown.

The first half of this thesis introduces a new class of time-dependent regime-switching (TDRS) SDEs, which can be shown to be stationary if we allow the parameters to be dependent on the elapsed time in a regime, instead of on the calendar time. The intuition behind our proposed class is that, although the process is time-dependent and non-stationary conditional on each realization of the hidden regime-switching (RS) process, the elapsed time in each state regime is a random time and implies stationarity of the TDRS SDE process under certain conditions. This type of intuition is motivated by the paper by Francq and Zakoïan (2001). These authors study multivariate RS ARMA models and have proved that local stationarity, i.e., stationarity within each regime, is not a necessary condition for obtaining the global stationarity for such models. In an SDE framework, we extend this idea by allowing the parameters in an RS SDE to have time-varying features.

The technique in the proof of the stationarity of our proposed model is similar to the one used by Mao and Yuan (2006), where the authors consider stationarity for a time-independent RS SDE. From a financial modeling perspective, our proposed model is a multi-factor one, and its randomness comes from a Brownian motion and a continuous-time RS process. It also incorporates the time-varying features of the parameters.

In Chapter 2, we propose a two-step estimation procedure in which Hamilton filtering (1990) is applied to estimate the parameters. The methodology is verified through simulation studies. We also apply our proposed model and estimation method to a US T-bill data set and find a dramatic change in the drift parameter during the early 1980s. The change corresponds to a monetary policy shift by the U.S. Federal Reserve.

The contributions of the first part of the thesis can be summarized as follows:

- We introduce a new class of time-dependent regime-switching (TDRS) models by incorporating a regime-switching feature and allowing the parameters to depend on the elapsed time after the regime has switched. Existence and uniqueness results are presented. A two-step maximum likelihood estimation procedure by use of an EM algorithm is proposed.
- A special class of TDRS models as an extension of the Vasicek model is proved to be stationary. Simulation and applications of the TDRS Vasicek model have been studied.

Part II: Nonparametric inference based on one single realization of a process

Parametric models enjoy the advantage of being analytically tractable and asymptotically efficient; however, they are restrictive and sensitive to deviations

from parametric specifications. On the other hand, nonparametric models are flexible and robust, but lead to less efficient inference procedures due to the infinite dimension of the parameter space. For any finite number of data observed in practice, we should not try to estimate the nonparametric function over the infinite dimensional space. Rather, we can only hope to estimate a finite number of parameters with certain accuracy. The difference between nonparametric inference and parametric inference then lies in the fact that, as we gather more observations of the model object, we can estimate more parameters without reduced accuracy of the chosen finite number of parameters. This difference is precisely the idea behind the sieve method, a parametric approach to studying nonparametric problems and first introduced in Grenander (1981).

The second half of the thesis explores non-parametric time-dependent SDEs and proposes an estimation method for projecting the time-dependent component of the drift similar in spirit to the sieve method. We emphasize that one key assumption of our study is that a continuous realization of the diffusion process is available. This assumption is crucial since it allows a complete identification of the diffusion parameter, and the likelihood ratio based on continuous realizations becomes ready for use. From a modeling perspective, we argue that with the high-frequency data available in the financial market and today's more powerful computer capacity, the continuity assumption of data is not too restrictive. The existing studies of sieve estimation of the drift parameter of diffusion processes, assuming continuous realizations, all assume that the number of paths of the diffusion process increases to infinity; see Prakasa Rao (2004), Stone and Huang (2003), and Nguyen and Pham (1982). By their argument, the dimension of sieve spaces should grow as more paths of the process become available. However, we rarely see multiple paths of economic and financial variables in practice. In contrast, we study an asymptotic scheme in which the trajectory length of the process increases to infinity, which to the best of our knowledge has not been studied using a sieve approach. The technical difficulty

is that the sieve spaces in our setting change with T.

The contributions of the second part of the thesis can be summarized as follows:

- For a class of SDEs, we propose a method to estimate a projection of the non-parametric time-dependent component of the drift. The method is similar in spirit to the sieve method. A continuous realization of the process is assumed, and we study the asymptotic scheme as the length of trajectory $T \to \infty$. We prove that our estimator is asymptotically consistent as long as the dimension of the sieve spaces increases at a controlled speed with the length of the observed trajectory of the process.

- A closed-form maximum likelihood estimator of a projection of time-dependent level function, for a time-dependent Vasicek model, into a finite dimensional space is obtained.

- A confidence interval for a projection of the time-dependent function in the drift of a class of SDEs is derived, and a hypothesis-testing procedure to determine the dimension of the parameter space is developed. Simulation and application results are presented. In finite samples, a basic parametric bootstrap method is proposed to correct the estimation bias of the mean-reverting speed parameter in the time-dependent Vasicek model.

The following is an outline of the rest of this thesis:

- Section 1.2 reviews background knowledge and relevant literature.

- Chapter 2 introduces the TDRS models and the estimation methodology for parameters in the models. The TDRS Vasicek model is introduced as a specific example of TDRS models. At the end, an empirical study with a CKLS (1992) data set is investigated.

- Chapter 3 studies theoretical properties of TDRS models, including existence and uniqueness results. The stationarity of the TDRS Vasicek model is proved.

- Chapter 4 studies a class of SDEs with nonparametric time-dependent drift, with Brownian motion as an important example. Our objective is to estimate a projection of the drift onto a finite dimensional space. An estimation method similar in spirit to the Sieve method is formulated, and the asymptotic consistency of the sieve-type estimators is proved. A confidence interval and hypothesis testing based on exact distribution are also developed.

- Chapter 5 studies a time-dependent Vasicek model with a non-parametric time-dependent level function. Our objective is to estimate a projection of the level onto a finite dimensional space. Due to the aliasing problem of the unknown mean-reverting speed parameter and the nonparametric level function, the inference problem is more challenging. We derive and prove the asymptotic consistency of a sequence of maximum likelihood estimators of a projection of the level function. An approximate confidence interval and hypothesis-testing procedure are developed. To improve the finite-sample performance of our inference method, a basic parametric bootstrap method is proposed to correct the bias for estimation of the mean-reverting speed parameter. At the end, an empirical study with a CKLS (1992) data set is investigated.

1.2 A Brief Review of Relevant Materials

1.2.1 Interest Rate Modeling

Concepts and Notations

Interest rate is a concept common to most people. When we deposit money in a bank, we all expect that the balance after one day will increase. The time value of money is the so-called interest. Suppose that we deposit $\$B_0$ in the bank today and receive $\$B_T$ after T years. Then the continuously compounded interest rate r, assumed to be constant for the moment, will satisfy:

$$B_0 e^{rT} = B_T \tag{1.11}$$

However, as implied by the observed short-term zero-coupon bond prices in the market, the interest rate is not constant. Therefore, we study $r(t)$ as a function of time. We also consider $\{r(t)\}$ as a random process, since future values of interest rates cannot be calculated with certainty. First, let us review some basic definitions and notations for interest rate (Brigo and Mercurio, 2001).

Bank account (money-market account). We define $B(t)$ to be the value of a bank account at time $t \geq 0$. We assume $B(0) = 1$ and that the bank account evolves according to the following differential equation:

$$dB(t) = r(t)B(t)dt, \quad B(0) = 1,$$

where $r(t)$ is a positive function of time. As a consequence,

$$B(t) = exp(\int_0^t r(s)ds).$$

Stochastic discount factor. The stochastic discount factor $D(t, T)$ between two time instants t and T is the amount at time t that is "equivalent" to one unit

of currency payable at time T and given by

$$D(t,T) = \frac{B(t)}{B(T)} = exp(-\int_t^T r(s)ds).$$

Zero-coupon bond. A $T-$maturity zero-coupon bond (pure discount bond) is a contract that guarantees its holder the payment of one unit of currency at time T, with no intermediate payments. The contract value at time $t < T$ is denoted by $P(t,T)$. Clearly, $P(T,T) = 1$ for all T.

Note that the stochastic discount factor is a random quantity that is the "equivalent amount of currency," while a zero-coupon bond is "the value of a contract." If $r(s)$ is a deterministic function of time, these two numbers are exactly equal. However, $r(t)$ is a stochastic process, and the price of a zero-coupon bond has to be known at time t. In fact, the price of a zero-coupon bond is the expectation of the stochastic discount factor under a risk-neutral measure.

Time to maturity. The time to maturity $T - t$ is the amount of time (in years) from the present time t to the maturity time $T > t$.

Continuously-compounded spot interest rate. The continuously-compounded spot interest rate prevailing at time t for the maturity T is denoted by $R(t,T)$ and is the constant rate at which an investment of $P(t,T)$ units of currency at time t accrues continuously to yield a unit amount of currency at maturity T. Written as a formula,

$$R(t,T) := -\frac{lnP(t,T)}{\tau(t,T)},$$

where $\tau(t,T)$ is the time difference between t and T. The continuously compounded interest rate is therefore a constant rate that is consistent with the zero-coupon bond prices in that

$$e^{R(t,T)\tau(t,T)}P(t,T) = 1.$$

The instantaneous interest rate (**short rate**) is defined as

$$r(t) = \lim_{T \to t^+} R(t, T).$$

Zero-coupon curve. The zero-coupon curve (yield curve or term structure of interest rates) at time t is the graph of the function

$$T \mapsto \begin{cases} L(t,T) & t < T \leq t + 1(\text{years}) \\ Y(t,T) & T > t + 1(\text{years}), \end{cases}$$

where $L(t, T)$ is the simply compounded interest rate, and $Y(t)$ is the annually compounded interest rate.

Simply compounded forward interest rate The simply-compounded forward rate prevailing at time t for the expiry $T > t$ and maturity $S > T$ is denoted by $F(t, T)$ and is defined by

$$F(t, T, S) := \frac{1}{\tau(T, S)}(\frac{P(t, T)}{P(t, S)} - 1).$$

It is the constant rate for the period (T, S) that is consistent with the observed market prices of bonds $P(t, T)$ and $P(t, S)$.

Instantaneous forward interest rate. The instantaneous forward rate prevailing at time t for the maturity $T > t$ is denoted by $f(t, T)$ and is defined as

$$f(t, T) := \lim_{S \to T^+} F(t, T, S) = -\frac{\partial ln P(t, T)}{\partial T}$$

so that we also have

$$P(t, T) = exp\ (-\int_t^T f(t, u)du).$$

Note that the forward interest rate is implied by the market prices of bonds and is hence a deterministic function of T at time t. However, if we fix maturity T as a future time, then $f(s, T)$ is a random quantity for $s > t$, with t being the current

time. If we model $f(s,T)_{s\geq t}$ as a random process for fixed T, we end up with the Heath-Jarrow-Morton (HJM) framework mentioned later.

Below, we briefly review the vast literature in statistical modeling for interest rate. Traditionally, the interest rate is modeled through continuous-time stochastic processes, mainly due to the ease of derivation using stochastic calculus tools. There are three types of interest rate models: 1) short rate models (equilibrium models), which study the stochastic evolution of instantaneous spot interest rate and produce the term structure of the interest rate as an output; 2) The HJM framework (no-arbitrage model), where we model instantaneous forward rates for fixed maturity date T and take the initial term structure of the interest rate as an input; 3) market models. The two most popular market models are the LIBOR market model (LFM) and the swap market model (LSM). These two market models are in agreement with the well-established formulas for pricing caps and swaptions in the market. Here we focus on the short rate models (equilibrium models), which are the earliest and quite richly represented in the literature. One of the earliest short rate models is the Vasicek model (1.1), a linear SDE with constant volatility, which hence can be solved explicitly. The bond and option prices can also be readily derived as a function of the current short rate $r(t)$ only.

While the Vasicek model is analytically tractable, it has a fundamental drawback in that the process can assume negative values. In view of this fact, Cox, Ingersoll, and Ross (1985) propose the CIR model (1.2). In it, $r(t)$ follows noncentral χ^2 distribution conditional on previous realization $r(s)$, and has asymptotic gamma distribution as t approaches ∞.

The Vasicek and CIR models are time-homogeneous models, meaning that the parameters in the drift and diffusion components of the postulated SDE (1.1) and (1.2) do not change over time. The derived bond prices for different maturities are perfectly correlated; thus, these models typically are not able to fit the observed term structure of interest rate. To address this issue, Hull and White (1994a)

assume that the instantaneous short-rate process evolves under the risk-neutral measure according to (1.5), where ν is chosen to exactly fit the term structure of interest rates being currently observed in the market. In fact, one can show that

$$\nu(t) = \frac{\partial f^M(0,t)}{\partial t} + af^M(0,t) + \frac{\sigma^2}{2a}(1 - e^{-2at}), \tag{1.12}$$

where $f^M(0,t)$ denotes the market instantaneous forward rate at time 0 for the maturity t, i.e.,

$$f^M(0,t) = -\frac{\partial lnP^M(0,t)}{\partial t}, \tag{1.13}$$

with $P^M(0,t)$ being the market discount factor for maturity t.

As an extension of the Vasicek and CIR models, Chan et al. propose the model (1.3) in 1992. Notice that (1.3) reduces to (1.1) when $\gamma = 0$, and (1.3) reduces to (1.2) when $\gamma = 1/2$. The authors estimate parameters in the process by using the Generalized Method of Moments (GMM) proposed in Hansen (1982). They also use hypothesis-testing techniques developed by Newey and West (1987) to evaluate a set of models nested under the model (1.3). The data set used by the authors consists of monthly observations of annual yields for one-month US Treasury bills (T-bills) from June 1964 to December 1989.

Many parametric models have been proposed to describe historical interest rate data, and several authors have devoted efforts to compare these models. In fact, there is no agreement on the linearity of the drift component in the stochastic differential equation employed for the short rate process. Some authors have tried to specify the short-rate SDE in a nonparametric form and test the parametric form as a null hypothesis. To name a few, Aït-Sahalia (1996b) has tested continuous-time univariate diffusion models by comparing the implied parametric density and the density estimated nonparametrically. The test statistic employed in the paper is a version of the Kullback-Leibler distance function. The author finds evidence of nonlinearity in the drift term and proposes a more general parametric univariate

diffusion model:

$$dr(t) = (\alpha_0 + \alpha_1 r(t) + \alpha_2 r^2(t) + \alpha_3/r(t))dt + \sqrt{\beta_0 + \beta_1 r(t) + \beta_2 r(t)^{\beta_3}} dW(t).$$

Stanton (1997) presents a technique for nonparametrically estimating continuous-time diffusion processes using a stochastic Taylor expansion. The author employs his technique in estimating the drift, diffusion and market price of interest risk. Stanton claims that there is evidence of substantial nonlinearity in the drift. The data set the author studies consists of daily values of secondary market yields on three-(six-)month T-bills between January 1965 and July 1995 for short-term interest rate (market price of interest rate), converted from discounts to annualized yields. Chapman and Pearson (2000) examine the nonparametric estimators proposed in Aït-Sahalia (1996) and Stanton (1997) by applying them to simulated data from a CIR process. The results suggest the same nonlinearity of the drift term as documented in Aït-Sahalia (1996) and Stanton (1997), yet the true drift is linear. The authors identify sources of bias in the estimators used by Aït-Sahalia (1996) and Stanton (1997) and claim that time series methods alone are not capable of producing evidence of nonlinearity in the drift. Takamizawa (2008) estimates nonlinear drift models of the short rate using both time series data (one-month Eurodollar deposit rate) and cross-sectional data (three- and six-month Eurodollar deposit rates). The author reports that nonlinear physical drift is not implied unless it is strongly affected by cross-sectional dimensions of the data, and that nonlinear risk-neutral drift is desirable to explain and predict observed patterns of yield spreads. It seems that the linearity of the drift component is still controversial.

Single factor models, such as the Vasicek and CIR models, assume that all bonds with different maturities are subject to the same source of random shock. However, this assumption is counter factual. Empirical evidence shows that zero-coupon bonds with different maturities are not perfectly correlated. To this end, multi-factor models have been studied, including Brennan and Schwartz (1979), Fong and Vasicek (1991), Longstaff and Schwartz (1992). Since the interest rate is

macroeconomic, it can be affected by government policies and economic conditions (expansion or recession). Regime-switching models are natural candidates to capture the potential structural breaks within the interest rate process. For example, Driffill et al. (2002) study the following class of continuous-time regime-switching CIR models in order to select the most appropriate parametrization:

$$-dr(t) = k(S(t))[\alpha(S(t)) - r(t)]dt + \sigma(S(t))\sqrt{r(t)}dZ(t),$$

where $Z(t)$ is a standard Brownian motion. The authors construct the likelihood function by exact discretization of the CIR process

$$r_{t+\triangle} = e^{-k(S(t))\triangle t}r(t) + (1 - e^{-k(S(t))\triangle t})\alpha(S(t)) + \sigma(S(t))\sqrt{r(t)}\sqrt{\frac{1 - e^{-2k(S(t))\triangle t}}{2k(S(t))}}\,\epsilon_{t+\triangle},$$

where $S(t)$ is the underlying Markov chain, which governs structural change of the model. It is concluded that the models preferred by goodness of fit criteria can be different from the ones with predictive power in terms of pricing. Recently Choi (2009) has studied a more general continuous-time regime-switching model:

$$dr(t) = (\alpha_{-1s_t}r(t)^{-1} + \alpha_{0s_t} + \alpha_{1s_t}r(t) + \alpha_{2s_t}r(t)^2 + \alpha_{3s_t}r(t)^3)dt + \beta_{s_t}r(t)^{\rho_{s_t}}dW(t), \tag{1.14}$$

where the regime index s_t follows a continuous-time first-order Markov chain with two states. The author applies a closed form approximation to the transition density function, as suggested by Aït-Sahalia (2002), and employs the recursive algorithm developed by Hamilton (1989) to obtain the MLE for the parameters in (1.14). The data set Choi has investigated is weekly three-month T bill rates from January 8, 1971 to December 26, 2003.

Pricing and Calibration

In the financial industry, interest rate modeling is often used for pricing and hedging interest rate derivatives. The parameters are typically determined by calibration to market data of financial products, such as bonds, caps, and swaptions.

As an example, we illustrate the pricing and calibration procedure by using the Vasicek model. Suppose that under the risk-neutral measure Q, the short-term interest rate follows an Ornstein-Uhlenbeck process:

$$dr(t) = k[\theta - r(t)]dt + \sigma dW(t), \quad r(0) = r_0,$$

where r_0, θ and σ are positive constants. Since this is a linear SDE, we can derive the explicit solution:

$$r(t) = r(s)e^{-k(t-s)} + \theta(1 - e^{-k(t-s)}) + \sigma \int_s^t e^{-k(t-u)} dW(u), \quad \text{for each } s \leq t.$$

The price of a zero-coupon discount bond $P(t,T)$, maturing at time T, can be derived by solving the following PDE (Vasicek 1977):

$$\frac{\partial P}{\partial t} + k(\theta - r)\frac{\partial P}{\partial r} + \frac{1}{2}\sigma^2 \frac{\partial^2 P}{\partial r^2} - rP = 0, \quad P(T,T) = 1,$$

or by taking expectation under the risk-neutral measure (cf. Brigo and Mercurio, 2001)

$$P(t,T) = E_t\{e^{-\int_t^T r(s)ds}\}.$$

We obtain

$$P(t,T) = A(t,T)e^{-B(t,T)r(t)},$$

where

$$\begin{aligned} A(t,T) &= exp\{(\theta - \frac{\sigma^2}{2k^2})[B(t,T) - T + t] - \frac{\sigma^2}{4k}B(t,T)^2\} \\ B(t,T) &= \frac{1}{k}[1 - e^{-k(T-t)}]. \end{aligned}$$

Once we have a theoretical pricing formula for the zero-coupon bonds, we may estimate parameters by calibrating the model to market data. Simply speaking, a good guess of parameters chooses the ones that are able to produce

prices that are as close to the observed data as possible. For example, suppose we observe market prices for zero-coupon bonds with different maturities $P^M(t,T_1), P^M(t,T_2), \cdots, P^M(t,T_n)$. We can then find parameters k, θ, and σ by minimizing the difference between the theoretical bond pricing formula and the observed market prices. One common criterion for the "difference" is the sum of square errors, i.e.,

$$(\hat{k}, \hat{\theta}, \hat{\sigma}) = \arg\min_{k,\theta,\sigma} \sum_{i=1}^{n} (P^M(t,T_i) - P(t,T_i))^2.$$

1.2.2 Regime-Switching Models

A regime-switching model can be defined as a pair of stochastic processes $(S(t), X(t))_{t\in\Lambda}$, where Λ is an index set for time, $X(t)$ is observable, and $\{S(t)\}$ is an unobserved stochastic process taking countably many values. When the pair of the processes are discrete in time, i.e. $\Lambda = \{0, 1, 2, \cdots\}$, a regime-switching model can encompass several of the well-known models proposed in the statistical literature, such as the probability mixture model and the Hidden Markov Model (HMM).

For a probability mixture model $(S(t), X(t))_{t\geq 0}$, $\{S(t)\}_{t\geq 0}$ is assumed to be a collection of independent discrete random variables with the same distribution. Suppose that $S(t) = i$ with probability p_i, $i = 1, 2, \cdots, k$; and that the distribution of $X(t)$ is completely determined by the value of $S(t)$. More explicitly, assume $X(t) = X_i$ when $S(t) = i$ and $X_i's$ are independent random variables with possibly different distributions. Therefore, all $X(t)'s$ have the same probability density function

$$\sum_{i=1}^{k} p_i f_{X_i}(x),$$

where $f_{X_i}(x)$ is the density function for X_i. Note that in the probability mixture model, $S(t)'s$ are independent random variables. In the case when $\{S(t)\}$ follows a

Markov chain and $X(t)$ depends on $S(t)$ only, $\{X(t), S(t)\}_{t\geq 0}$ is a Hidden Markov Model (HMM). By definition, an HMM (Figure 1.2) is a statistical model in which the system being modeled is a Markov process with unobserved states. It can be described by a pair $(S_k, X_k)_{k\geq 1}$ such that

$$\begin{aligned} \{S_k\} &: \quad \textit{a Markov chain with a finite number of states } N, \textit{ unobservable} \\ X_k &: \quad \textit{a r.v. with distribution associated with each state of } S_k, \textit{observable.} \end{aligned}$$

Such models can be dated back at least to Baum and Petrie (1966). Their applications include cryptanalysis, speech recognition, machine translation, gene prediction, and partial discharge (Satish and Gurura, 2003). Although S_k is not directly observable, the observed value of X_k is generated conditional on the state of S_k. Therefore, we are able to draw statistical inference for S_k through realizations of X_k. The parameters in the HMM include the transition probabilities of S_k and other parameters governing the distribution of S_k and X_k. The consistency and asymptotic normality of the maximum likelihood estimator of the parameters of HMMs have been considered by numerous authors, including Leroux (1992), Bickel et al. (1998), and Le Gland and Mevel (2000).

Regime-switching regressions were first introduced by Goldfeld and Quandt (1973). In their paper, the authors study the demand and supply functions of housing markets. However, the popularity of regime-switching models in recent decades is due largely to the seminal paper by Hamilton (1989). This author proposed the following regime-switching autoregressive model for GNP (Gross National Produce) data of the US:

$$(x_t - \mu_{s_t}) = \phi_1(x_{t-1} - \mu_{s_{t-1}}) + \phi_2(x_{t-2} - \mu_{s_{t-2}}) + \cdots + \phi_r(x_{t-r} - \mu_{s_{t-r}}) + \epsilon_t, \tag{1.15}$$

where $\{\epsilon_t\}_{t\geq 1}$ are normal innovations and $\{S(t)\}$ is an unobservable Markov chain with two states. In the paper, the author derives an iterative algorithm similar in

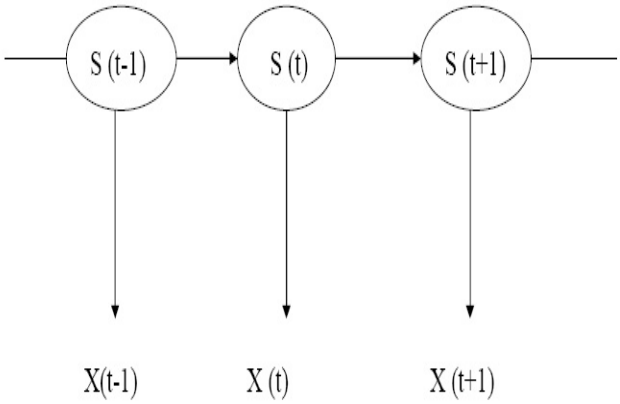


Figure 1.2: Hidden Markov Model

spirit to the Kalman Filter. Moreover, in order to improve computational efficiency and robustness for the maximum likelihood estimation, the author introduces an EM algorithm to estimate the model (1.15) in another paper published in 1990 and derives analytic derivatives of the sample log-likelihood function. According to Hamilton (1990), the advantages of EM algorithms are their robustness and saving of computational time, in comparison to maximizing the likelihood function numerically over an often ill-behaved likelihood surface with respect to a large number of unknown parameters.

Apart from the above-mentioned econometrics applications, regime-switching models have also been applied in many other areas. For example, in the field of actuarial science, Hardy (2001) introduces a regime-switching model for equities. The author employs a regime-switching lognormal model to study monthly S&P 500 and TSX 300 indices and develops a European option-pricing formula by conditioning on the number of months spent in each regime. In the context of interest rate modeling, Gray (1996) proposes a univariate regime-switching GARCH model. Later, Ang and Bekaert (2002) study a multivariate regime-switching model by in-

corporating international short-rates and term spreads. Bansal and Zhou (2002) develop a term structure model for which the short rate and market price of risks are both subject to regime shifts. The aforementioned short rate models assume that the short term interest rate follows a discrete-time process. Continuous-time regime-switching short rate models have been studied by Naik and Lee (1997), Drifill et al. (2004), Wirjanto (2006), and Choi (2009).

The following is a brief overview of statistical inferences for regime-switching models. Krishnamurthy and Rydén (1998) study the following process:

$$X_n = g(X_{n-1}, \cdots, X_{n-d}, e_n; \theta_{S_n}(\Phi)), \tag{1.16}$$

where $\{g(\cdot, \theta) : \theta \in \Theta\}$ is a family of real-valued functions defined on $\mathbb{R}^{d+1}$, indexed by a parameter $\theta \in \Theta$; $d > 0$ is a fixed and known integer; Θ is a Euclidean space; S_n is a finite state Markov chain, and $\{e_n\}$ is an independent and identically distributed (i.i.d.) sequence of innovations. Let $Z_k = (S_k, X_k, X_{k-1}, \cdots, X_{k-d+1})$. Assuming the existence of an ergodic stationary solution Z_k, the authors have proved the consistency of the MLE under certain technical conditions by following an approach used in Leroux (1992). In the case when $g(\cdot)$ is linear, a sufficient condition for the existence of a unique strictly stationary solution has been given by Brandt (1986), Karlsen (1990), and Bougerol and Picard (1992). Another class of models has been investigated by Francq and Roussignol (1998), who examine a Markov-switching autoregressive time series model :

$$X(t) = F(X_{t-1}, S(t), \lambda) + G(\eta(t), S(t), \lambda), \quad \forall t \geq 1,$$

where $(\eta(t))$ is a sequence of independent and identically distributed multivariate random vectors; λ is an unknown parameter belonging to an open subset Θ of $\mathbb{R}^d$; $\{S(t)\}$ is independent of $\{\eta(t)\}$, and $F(\cdot)$ and $G(\cdot)$ are measurable functions. The authors give conditions for the existence of an ergodic stationary solution to the model and also prove the consistency of the maximum likelihood estimator. Note

that this model is in a switching regression form, which is not as general as model (1.16). However, the results are stronger, and the techniques employed are different. To address the stationarity issues of regime-switching models, Francq and Zakoïan (2001) consider multivariate ARMA models with random coefficients:

$$X(t) = c(S(t)) + \sum_{i=1}^{p} a_i(S(t))X_{t-i} + \epsilon_t + \sum_{j=1}^{q} b_j(S(t))\epsilon_{t-j}, \tag{1.17}$$

where $X(t)$ is a random vector with values in $\mathbb{R}^K$; and $(S(t))$ is an irreducible, aperiodic, Markov chain with finite state-space. The authors have derived a sufficient condition for the existence of a strictly stationary solution. Moreover, they give a sufficient condition, which is also necessary when $c(\cdot) = 0$, for the existence of a second-order stationary solution. The authors present examples showing that local stationarity, i.e., stationarity within each regime, is neither necessary nor sufficient to ensure global stationarity. Douc, Moulines, and Rydén (2004) have proved the consistency and asymptotic normality of the maximum likelihood estimator where the hidden state space is compact. Moreover, the authors have extended their results to non-stationary AR models with Markovian regimes.

All the regime-switching models mentioned previously are discrete-time stochastic processes. Recently, the stochastic differential equations with Markovian switching (regime-switching) (2.2) have also received a great deal of attention (Mao and Yuan, 2006). Applications of SDEs with Markovian switching include population dynamics, financial modeling, stochastic stabilization, and stochastic neural networks (Mao and Yuan, 2006).

1.2.3 Computational Methods for Finding the Maximum

The maximum likelihood estimator (MLE) is defined as the solution to the following equation:

$$\hat{\theta} = \arg\max_{\theta} L(\theta; x),$$

where $L(\theta, x)$ is the likelihood function. The advantages of MLE include its invariance properties and asymptotic efficiency (lower variance). However, when data is incomplete, which is the case for continuous-time regime-switching SDEs, a closed-form likelihood function is generally not available. In this case, the EM algorithm and simulated likelihood method are known to be powerful computational tools for dealing with incomplete data problems.

The EM Algorithm

The EM algorithm was explained and given its name in Dempster, Laird, and Rubin (1977), who developed a general theory. The advantages of the EM algorithm include its ease of implementation (simplicity), guaranteed increment of likelihood function for iterated parameter values (monotonicity), and robustness to the initial parameter values (stability). As a computational algorithm, the EM algorithm does not automatically produce a covariance matrix and suffers from slow convergence. However, some work has been done to attenuate these issues. (Mclachlan and Krishnan, 2008).

Before we formulate the method, we provide an intuitive explanation of why the EM algorithm works. Suppose X represents a random variable with density function $f_\theta(x)$. Let $Y = Y(X)$ be a function of X with density function $g_\theta(y)$. Suppose further that $f_\theta(x)$ and $g_\theta(y)$ are regular models. We then have the key relation

$$\frac{\partial}{\partial\theta}\log g_\theta(y) = E[S(\theta; X)|Y = y],$$

where $S(\theta; x) = \dfrac{\partial \log f_\theta(x)}{\partial\theta}$ is the score function. To show this relation, let us

consider a measurable set $B \in \sigma(\mathbb{R})$. Then,

$$\begin{aligned}
\int_B \frac{\partial \log g_\theta(y)}{\partial \theta} g_\theta(y) dy &= \int_B \frac{\partial g_\theta(y)}{\partial \theta} dy = \frac{\partial}{\partial \theta} P_\theta(Y \in B) = \frac{\partial}{\partial \theta} P_\theta(Y(X) \in B) \\
&= \frac{\partial}{\partial \theta} \int_{x \in Y^{-1}(B)} f(\theta; x) dx = \int_{x \in Y^{-1}(B)} \frac{\partial f(\theta; x)}{\partial \theta} dx \\
&= \int_{x \in Y^{-1}(B)} S(\theta; x) f(\theta; x) dx = E[S(\theta; X) I_{\{Y(X) \in B\}}] \\
&= E[E[I_{\{Y(X) \in B\}} S(\theta; X) | Y]] = E[I_{\{Y(X) \in B\}} E[S(\theta; X) | Y]] \\
&= \int_{y \in B} E[S(\theta; X) | Y = y] g_\theta(y) dy.
\end{aligned}$$

Since B is arbitrary, we have

$$\frac{\partial}{\partial \theta} \log g_\theta(y) g_\theta(y) = E[S(\theta; X) | Y = y] g_\theta(y).$$

Therefore,

$$\frac{\partial}{\partial \theta} \log g_\theta(y) = E[S(\theta; X) | Y = y].$$

Based on the above argument, the EM algorithm can be applied in two steps:

1. *E (Expectation) step*:

$$E_{\theta^{(l)}}[\log f(\theta, X) | Y = y], \tag{1.18}$$

   where we compute the conditional log-likelihood function of the complete data given the observed data and starting parameter $\theta^{(l)}$.

2. *M (Maximization) step*:

$$E_{\theta^{(l)}}(S(\theta^{(l+1)}; X) | Y = y) = 0,$$

   where we maximize the function (1.18) from the E step by applying the first-order condition.

After we obtain $\theta^{(l+1)}$, we repeat our E step and M step until the sequence converges to a real number, which is the maximum likelihood estimate.

Maximum Simulated Likelihood

When the exact likelihood function is not known in a closed form, we can consider using a simulated likelihood, which is an approximation of the true likelihood function obtained through simulations. Intuitively, simulated likelihood means expressing the likelihood function as an expectation of a random variable and then employing Monte Carlo methods to obtain the expected value of this random variable. More explicitly, let $Y = (Y_1, Y_2, \cdots, Y_n)$ be a random vector with joint probability density function $f_\theta(y_1, y_2, \cdots, y_n)$ and $X = (X_1, X_2, \cdots, X_m)$ be another random vector with joint probability density function $g_\theta(x_1, x_2, \cdots, x_m)$, where θ is a vector of unknown parameters. We then have

$$E[f_\theta(y_1, \cdots, y_n | X_1, \cdots, X_m)] = f_\theta(y_1, \cdots, y_n),$$

where $f_\theta(\cdot|\cdot)$ is the conditional likelihood function. In fact,

$$\begin{aligned} E[f_\theta(y_1, \cdots, y_n | X_1, \cdots, X_m)] &= \int f_\theta(y_1, \cdots, y_n | x_1, \cdots, x_m) g_\theta(x_1, \cdots, x_m) dx_1 \cdots dx_m \\ &= \int f_\theta(x_1, \cdots, x_m, y_1, \cdots, y_n) dx_1 \cdots dx_m \\ &= f_\theta(y_1, \cdots, y_n). \end{aligned}$$

Let $(X^1, \cdots, X^j, \cdots, X^K)$ be independent simulations of the random vector $X = (X_1, \cdots, X_m)$. Then, by WLLN,

$$\tilde{f}_\theta(y_1, \cdots, y_n) := \frac{\sum_{i=1}^K f_\theta(y_1, \cdots, y_n | X^i)}{K} \xrightarrow{P} f_\theta(y_1, \cdots, y_n).$$

Moreover, we have

$$E[\tilde{f}_\theta(y_1, \cdots, y_n)] = f_\theta(y_1, \cdots, y_n).$$

Furthermore, we define the log-likelihood function and MLE:

$$\begin{aligned} l(\theta) &= ln f_\theta(y_1, \cdots, y_n) \\ \tilde{l}(\theta) &= ln \tilde{f}_\theta(y_1, \cdots, y_n) \\ \theta_{MLE} &= \arg\max_\theta \ l(\theta) \\ \tilde{\theta}_{MSL} &= \arg\max_\theta \ \tilde{l}(\theta) \end{aligned}$$

Materials in the following papers are closely related to the application of simulated likelihood methods in my research work. Pedersen (1995a, 1995b) introduced a simulated maximum likelihood (SML) method for stochastic differential equations based on discrete observations when the likelihood function is unknown. The author approximates the log-likelihood function by simulating intermediate points in between observations. During the last step, Pedersen uses normal distribution to write out the simulated likelihood function. In the context of model selection by hypothesis test, Durham (2003) evaluated a sequence of nested continuous-time models for short-term interest rate and found that allowing for additional flexibility beyond a constant term in the drift provides little benefit. The analysis he uses involves likelihood-based information criteria (Akaike Information Criterion, Schwarz Criterion, and the likelihood ratio test), through a simulated maximum likelihood procedure proposed in Durham and Gallant (2002). To speed up the algorithm, Sørensen (2003) proposed a k-th order approximation to the true likelihood function for a class of discretely observed continuous-time stochastic volatility models:

$$\begin{aligned} dX(t) &= \xi(V(t))dt + \sqrt{V(t)}dW(t) \\ dV(t) &= b(V(t), \theta)dt + \sigma(V(t), \theta)dW(t). \end{aligned}$$

The author approximates the likelihood function by pretending that the observed values follow a k-th order Markov process. Later on, the k-th order likelihood function is approximated again by simulation of the latent process through the Milstein scheme. The author also proves the consistency and asymptotic normality

of MLE based on a k-th order approximate likelihood for any fixed k and $\triangle$ (sampling frequency). Finally, the author illustrates the method through simulation of a stochastic volatility model considered by Hull and White (1987) and Heston (1993). However, the drift and diffusion functions of the stochastic volatility model do not depend on $X(t)$ itself and, hence, we know the conditional distribution of $X(t)$ given the latent process $V(t)$. Note also that the discretized version of the studied model is a Hidden Markov model with a continuous, unbounded state space of the underlying Markov chain.

Maximum Simulated Likelihood is relatively easy to implement with mathematical software. However, doing so is not always feasible in practice due to computational budget restrictions. One computational cost is related to the number of simulations required to obtain a precise likelihood value. Typically, a path of the latent process must be run, and a huge number of simulations may be necessary to calculate a precise likelihood value for the observed process. Another source of computational cost is optimization of the likelihood surface to obtain the MLE. As documented by Sørensen (2003), it is crucial to use the same random numbers to calculate the likelihood of different values of θ. Otherwise, obtaining a stable and reliable estimate requires many more simulations.

Direct Likelihood

Suppose $(S(t), X(t))_{t=1,\cdots,n}$ is a regime-switching process, where $\{X(t)\}$ is observable and $\{S(t)\}$ is a latent Markov chain. Francq and Zakoïan (2001) study a two-regime AR(1) model:

$$X(t) = c(S(t)) + a(S(t))X_{t-1} + \sigma(S(t))\eta(t),$$

where $\{\eta(t), t = 1, 2, \cdots, n\}$ are i.i.d. N(0,1) random variables. Let $\{p(i,j) : i,j = 1,2\}$ be transition probabilities and $\theta = \{p(1,1), p(2,1), c(1), c(2), a(1), a(2), \sigma(1), \sigma(2)\}$ be all the parameters. The

likelihood function is directly calculated by matrix multiplications, which can be written as

$$L_\theta(X_1, \cdots, X_n) \tag{1.19}$$

$$= \sum_{s_1,\cdots,s_n} \pi(s_1) \prod_{i=2}^{n} p(s_{i-1}, s_i) \prod_{i=2}^{n} f_{s_i}(X_{i-1}, X_i). \tag{1.20}$$

where

$$f_{s_i}(X_{i-1}, X_i) = \frac{1}{\sqrt{2\pi\sigma(s_i)}} \exp\left(-\frac{[X_i - c(s_i) - a(i)X_{i-1}]^2}{2\sigma(s_i)^2}\right).$$

Let $\mathbf{1} = (1, 2)' \in R^2$, $p(X_1) = (\pi(1)f_1(X(0), X_1), \pi(2)f_2(X(0), X_1))' \in R^2$ and

$$A_\theta(X_{i-1}, X_i) = \begin{pmatrix} p(1,1)f_1(X_{i-1}, X_i) & p(2,1)f_1(X_{i-1}, X_i) \\ p(1,2)f_2(X_{i-1}, X_i) & p(2,2)f_2(X_{i-1}, X_i) \end{pmatrix}.$$

Then it is easy to verify that

$$L_\lambda(X_1, \cdots, X_n) = \mathbf{1}'[\prod_{i=2}^{n} A_\lambda(X_{n+1-i}, X_{n+2-i})]p(X_1), \tag{1.21}$$

where

$$[\prod_{i=2}^{k} A_\lambda(X_{k+1-i}, X_{k+2-i})] = A_\lambda(X_{k-1}, X_k)[\prod_{i=2}^{k-1} A_\lambda(X_{k-i}, X_{k+1-i})]. \tag{1.22}$$

1.2.4 The Sieve Method

In statistical practice, parametric models, for which the model is specified with a finite number of parameters, are often found to be easy to estimate. However, they are rather restrictive and have potential risks of misspecification of the relationship between variables. On the other hand, nonparametric models, for which the dimension of parameter space is infinite, are more flexible and robust, but the optimization problem for finding parameter estimates might have no solution or

the resulting estimator might not be consistent. For example, the maximum likelihood estimator of a probability density function over $L^1(\mathbb{R})$ based on observation of i.i.d random variables $x_1, x_2, \cdots, x_n$ is not attainable. Therefore, nonparametric estimation methods, such as the kernel method, local linear regression and sieve methods, have been developed to tackle this problem.

Sieve estimators are a class of nonparametric estimators that use progressively more complex models to estimate an unknown high-dimensional function as more data becomes available, with the aim of asymptotically reducing error towards zero as the amount of data increases. This method was introduced to statistics in Grenander (1981). To help the reader understand the name "sieve", we give the following definition from Grenander (1981):

> A sieve, usually denoted by $\mathcal{S}(\mu)$, is a family of subsets of Θ (parameter space) indexed by a positive parameter μ, the *mesh* size. For any $\mu > 0$, $\mathcal{S}(\mu)$ shall be sufficiently restricted to make a ML(maximum likelihood) solution exist. On the other hand, as the mesh size tends to zero, the set $\mathcal{S}(\mu)$ shall be sufficiently rich to allow the ML solution to converge to any $\theta \in \Theta$.

The sieve method consists of two key ingredients: a criterion function and sieve parameter spaces (a sequence of approximating spaces). A criterion function is a function from the parameter space to real numbers, which is uniquely maximized at the true parameter $\theta_0 \in \Theta$. For the maximum likelihood estimation, the criterion function is the likelihood function; for Generalized Method of Moments (GMM), the criterion function is of the quadratic form $g'(\theta)Wg(\theta)$, where $g(\theta)$ is a vector of unconditional moment conditions and W is a possibly random weighting matrix. The choice of the sieve parameter space depends on how well the sieve spaces approximate Θ and how easily we can compute the estimator over the sieve spaces. For example, the sieves or approximating spaces can be constructed using linear

spans of power series, Fourier series, splines or many other basis functions. To ensure consistency of the method, we require that the complexity of sieves increases with the sample size so that, in the limit, the sieves are dense in the original parameter space. Since these approximating spaces can often be characterized by a finite number of "parameters", a nonparametric or semi-parametric estimation problem is often reduced to a parametric one when the sieve method is implemented. However, to obtain the desired theoretical properties of the estimator, the number of parameters must increase at a controlled speed with the sample size. It is this feature that gives the sieve method its added flexibility and robustness over classical parametric methods, which assume fixed, finite-dimensional parameter spaces.

There are numerous applications of the sieve method in the econometrics literature. For an excellent and detailed review of applications of the sieve method in non-parametric and semi-nonparametric modeling in econometrics, we refer the reader to Chen (2007). Here we give a brief summary of the applications and existing results in the literature. In the microeconometrics context, Newey and Ridder (2005) use power series and splines in the two-step efficient estimation of the average treatment effect models. Blundell et al. (2007) consider a profile sieve minimum distance (MD) procedure to estimate shape-invariant Engel curves with nonparametric endogenous expenditure. Chen et al. (2006) study the sieve MLE of semi-nonparametric multivariate copula models. Bierens (2008) and Bierens and Carvalho (2007) use orthonormal Legendre polynomials to model semi-nonparametrically the unobserved heterogeneity distribution of interval-censored mixed proportional hazard models and bivariate mixed proportional hazard models, respectively. In a time series econometric context, Engle et al. (1986) forecast electricity demand using a partially linear spline regression. Gallant and Tauchen (1996, 2004) have proposed combining a Hermite polynomial sieve and simulated method of moments to effectively solve many complicated asset-pricing models with latent factors, and their methods have been widely applied in empirical finance.

Engle and Rangel (2004) propose a new Spline GARCH model for measuring unconditional volatility and have applied it to equity markets for 50 countries for up to 50 years of daily data.

Large sample theory for the sieve method not only accounts for the approximation errors, which arise because we replace the original parameter space with the simpler sieve space, but also controls the complexity of sieve parameter spaces, which increases with the sample size. Consequently, the large sample properties of the sieve method are in general difficult to derive. For an infinite-dimensional, possibly noncompact parameter space Θ, Geman and Hwang (1982) obtain the consistency of sieve MLE with i.i.d. data. White and Wooldridge (1991) obtain the consistency of sieve extremum estimates with dependent and heterogeneous data. Newey and Powell (2003) and Chernozhukov, Imbens and Newey (2007) establish the consistency of sieve MD estimates. Bierens (2011) proves the consistency and $\sqrt{N}$ asymptotic normality of a sieve estimator of an unknown density function when a semi-nonparametric (SNP) discrete choice index model is considered.

There are many results on convergence rates of sieve M-estimators of unknown functions. For i.i.d. data, Van de Geer (1995) obtain the rate for sieve LS regression. Shen and Wong (1994) and Birgé and Massart (1998) derive the rates for general sieve M-estimation. Van de Geer (1993) and Wong and Shen (1995) obtain the rates for sieve MLE. For time series data, Chen and Shen (1998) derive the rate for sieve M-estimation of stationary beta-mixing models. To date, we have a relatively complete theory of the rates of convergence for sieve M-estimators. The existing asymptotic distribution theory results are mostly for series estimators of densities and the LS regression functions. Asymptotic normality of the series LS estimators has been studied in Andrews (1991b), Gallant and Souza (1991), Newey (1994b, 1997), Zhou et al. (1998), and Huang (2003). Stone (1990) and Strawderman and Tsiatis (1996) have given asymptotic normality results for polynomial spline estimators in the context of density estimation and hazard estimation, respectively.

There are also specification testing results that use the sieve method. For example, Hong and White (1995) test a parametric regression model using series LS estimators; Hart (1997) presents many consistent tests using series estimators; Stinchcombe and White (1998) test a parametric conditional moment restriction using neural network sieves, and Li et al. (2003) test semiparametric/nonparametric regression models using spline series estimators. Song (2005) proposes consistent tests of semi-nonparametric regression models via conditional martingale transforms.

The above results are all in econometrics literature, most of which has been covered in Chen (2007). In the following, we list some results in the diffusion process framework. Statistical inference for stochastic processes are of considerable importance and have been extensively studied in the past three decades; see Prakasa Rao (1999) for a list of many references. Grenander (1981) applies his method of sieves to estimate the drift function in a Brownian motion:

$$dX(t) = \theta(t)dt + dW(t).$$

Geman and Hwang (1982) show how to choose the dimension of sieve spaces so that the consistency of sieve estimators can be secured. Nguyen and Pham (1982) study the following non-stationary linear diffusion process:

$$dX(t) = \theta(t)X(t)dt + dW(t).$$

The authors use an increasing sequence of finite dimensional subspaces of the parameter space as the natural sieves on which a maximum likelihood estimation method is used. They prove that the sequence of restricted maximum likelihood estimators is consistent and asymptotically normal when the dimension of the sieves tends to infinity not too fast as the number of independent continuous realizations of the process increases. Mckeague (1986) uses a sieve method to estimate time-dependent covariates for the following semi-martingale regression model:

$$X(t) = X(0) + \int_0^t \alpha(s)Y(s)ds + M(t), t \in [0, 1],$$

where Y is a covariate process and M is a square integrable martingale. Beder (1987) presents a sieve estimator of the mean function $m(t)$ of a general Gaussian process. The author proves that his estimator is asymptotically unbiased and consistent at each t. Stone and Huang (2003) study the following model

$$dY(t) = \eta(t, X(t))dt + \sigma(t)dW(t), 0 \leq t \leq \tau,$$

where $0 < \tau < \infty$. It is assumed that the diffusion coefficient $\sigma(t)$ at time t is a known random function of time. Based on many continuous realizations of the process, the authors construct a sequence of restricted maximum likelihood estimators over spaces spanned by polynomial splines. They obtain rates of convergence of spline estimates for both fixed and free knot spline estimates. Prakasa Rao (2004) studies

$$dX(t) = \theta(t)X(t)dt + dW(t)^H,$$

where $\{W(t)^H\}$ is a fractional Brownian motion. The author studies the maximum likelihood estimator by using the sieve method in an approach similar to Pham and Nguyen's (1982). All these papers assume that independent continuous realizations of the process are available and prove asymptotic results as the number of realizations n tends to infinity.

There are also some applications of the sieve method when the data is discretely observed. Genon-Catalot et al. (1992) estimate the diffusion coefficient using the sieve method based on discrete data by assuming that the sampling interval converges to zero. Darolles and Couriéroux (2000) have studied inference on continuous-time processes from discrete data. Their approach consists of truncating the initial process to improve the estimation of the eigenfunctions. For the following model,

$$dX(t) = \mu(X(t))dt + \sigma(X(t))dW(t),$$

the authors propose a a modification of the sieve estimation method based on the infinitesimal generator and have derived estimators for the drift and volatility.

Chapter 2

A New Class of Time-Dependent Regime-Switching Models

2.1 Introduction

Ever since Hamilton published his seminal paper in 1989, regime-switching models have received considerable attention from researchers in the financial and economic areas. For example, many papers document evidence of regime changes in the evolution of interest rates (Hamilton, 1988; Garcia and Perron, 1996; Ang and Bakaert, 2002). However, most existing studies focus on discrete-time models, such as AR, ARMA, and ARCH. Only a few researchers have investigated regime-switching continuous-time diffusion processes (Naik and Lee, 1997; Driffill et al., 2004; Choi, 2009). When applied to interest rates, regime-switching models require that the model does not change within a particular regime. However, it is reasonable to expect that in some situations the parameters may be time-dependent within the same regime. We can then interpret a shift of regime as a response to a significant change in economic conditions. Such models would change within a

regime as a typical evolution in time. For example, such dynamics would capture better situations where financial data respond more abruptly at the beginning of an economic crisis than they do over time. Thus it is reasonable to introduce time dependency into a regime-switching diffusion process. As a result, we propose the following class of Time-Dependent Regime-Switching (TDRS) models:

$$dX(t) \quad = \quad f(X(t), S(t), \beta(t))dt + g(X(t), S(t), \beta(t))dW(t), \tag{2.1}$$

where

- $S(t)$ is a continuous-time Markov chain with finite state space $\mathbb{S} = \{1, 2, \cdots, N\}$ (whose states we will call regimes), which we assume to be unobservable and independent of the Brownian motion $\{W(t)\}$.
- $\beta(t)$ is the time that has elapsed in the current regime, which we also assume to be unobservable.
- $X(t)$ is a stochastic process whose evolution depends on $S(t)$, which we assume to be observable.

A similar type of model has been introduced in Mao and Yuan (2006):

$$dx(t) = f(t, x(t), S(t))dt + g(t, x(t), S(t))dW(t). \tag{2.2}$$

The main difference between this model and our proposed model is that the former depends on calendar time, while the latter depends on the time spent in the current regime. We would like to point out that most of the work done so far on the model (2.2) is related to its probabilistic aspects, such as the existence of solutions, stability, and boundedness. Few results are available regarding statistical issues, such as the consistency and asymptotic normality of (non)parametric estimation procedures.

Redekop and Wirjanto (2010) consider a functional of the path of a two-state Markovian switching diffusion process, which is the integral of squared volatility.

The authors derive the distribution function and moment generating function for the time spent in the high-volatility state on a fixed time interval [0,T].

Another relevant work in the statistical literature is by Davis (1993). The author has proposed a Piecewise Deterministic Process (PDP), wherein a sequence of events occurs at fixed or random times $T_1 < T_2 < T_3 \cdots$, and the process evolves as a deterministic function of time between event occurrences. Loosely speaking, a PDP is a mixture of deterministic motion and random jumps, and jump times follow a Poisson process. Note that PDP is a stochastic model that does not lie within the SDE framework, since it does not involve Brownian motion. Therefore, PDP does not fit in with the current finance literature very well. The layout of the rest of this chapter is as follows:

- Section 2.2 introduces the general form of TDRS models and provides one important example: the TDRS Vasicek model.
- Section 2.3 describes the methodology for parameter estimation. An Euler discretization of the process and a truncation of an infinite order Markov chain are proposed.
- Section 2.4 describes in detail the MLE based on a filtering method proposed in Hamilton (1990). Using this approach, we can infer the hidden regime based on the entire history of observed data and then apply the EM algorithm to find the MLE.
- Section 2.5 studies the TDRS Vasicek model. Simulation and application results for the model are also presented.
- Section 2.6 draws concluding remarks.

2.2 General Time-Dependent Regime-Switching (TDRS) Models

In Mao and Yuan (2006), the following time-dependent regime-switching SDE has been studied:

$$dX(t) = f(X(t), t, S(t))dt + g(X(t), t, S(t))dW(t), \tag{2.3}$$

where $S(t)$ is a continuous-time Markov chain with a finite number of states N. In this model, the components of the drift and diffusion parameters change according to calendar time in a deterministic way. Generally speaking, the solution to (2.3) is not stationary, which makes statistical inference on (2.3) quite challenging.

In the context of regime-switching models, it is reasonable to consider the situations in which the dependence of parameters on time is the same within a given regime so that the regime-switching process can assume a stationary distribution in the long run. In the following we propose a new class of time-dependent regime-switching model where the components of drift and diffusion parameters change according to the time that has elapsed in the current regime. We call this new class of models time-dependent regime-switching (TDRS) models and use this name for the rest of this thesis. A general form of the proposed model is given in (2.1), where $\beta(t) = t - \tau_t$, and τ_t is the last transition time of the underlying Markov chain. Note that $\beta(t)$ denotes the time that has elapsed in the current regime of the process $\{S(t)\}_{t\geq 0}$, and hence the value of $\beta(t)$ is deterministic given the realization of $\{S(t)\}_{t\geq 0}$.

Remark 2.1. *Note that in model (2.1), $(X(t), S(t), \beta(t))$ is a three-dimensional Markov process. However, $(X(t), S(t))$ is not a Markov process because the future distribution of $X(t)$ depends on $\beta(t)$ as well.*

The main example I will use to illustrate the general model (2.1) is

$$dX(t) = a(\theta(S(t), \beta(t)) - X(t))dt + \sigma dW(t), \tag{2.4}$$

where $\theta(S(t), \beta(t))$ is a time-dependent parameter in the drift term. We will refer to (2.4) as the TDRS General Vasicek model. In this model we can choose, for example, the following form of the time-dependent component:

$$\theta(S(t), \beta(t)) = \theta_3 + (\mu(S(t)) - \theta_3)e^{c\beta(t)}, \text{with } \theta_1 < \theta_3 < \theta_2, \tag{2.5}$$

where the components are interpreted in the following way:

- $\{S(t)\}$ is a continuous-time Markov chain with two regimes $\{1, 2\}$. State 1 is called the low regime, and state 2 is called the high regime.
- μ is a function of the current state defined as follows:
  - If $S_t = 1$, then $\mu(S_t) = \theta_1$.
  - If $S_t = 2$, then $\mu(S_t) = \theta_2$.
- c is a negative real number, indicating the speed of convergence to an equilibrium.

If the state process jumps to state 1, then θ function starts to increase exponentially; if the process jumps to state 2, then θ decays exponentially. Later we will provide an example with particular parameterization. In the remainder of this thesis, we will refer to the model (2.4) with mean-reversion level function given in (2.5) as the TDRS Vasicek model.

The TDRS General Vasicek model and the TDRS Vasicek model are motivated by the well-known Vasicek model, which has received a lot of attention in the finance literature due to both its analytic tractability and guaranteed stability. They are classified as mean-reverting models due to the specification in the drift coefficient.

When taking into account the regime shifts (economic expansion or recession), it is reasonable to expect that the level of interest rate reverts to an equilibrium value (the long-run average) if that rate stays in the same regime for a long time. When economic conditions change abruptly, such as from expansion to contraction, the process may switch to another regime abruptly. Therefore, the TDRS Vasicek model can be seen as a natural generalization of the Vasicek model. This intuition is consistent with common practice where calibrated models are time-dependent.

Some basic features of the TDRS Vasicek model include:

- every time the regime switches, the level $\theta(t, S(t))$ is brought back to common prescribed values θ_1 or θ_2.
- the process can jump to the levels θ_1 or θ_2 and reverts back to θ_3 at the rate specified by the parameter c.
- when $c = 0$, then the TDRS Vasicek reduces to a time-homogeneous regime-switching Vasicek model.
- when $c < 0$, if the process stays in the same regime for a long time, then the level $\theta(t, S(t))$ converges to a constant θ_3, and the equilibrium model is

$$dX(t) = a(\theta_3 - X(t))dt + \sigma dW(t),$$

  which is the Vasicek model.

Figure 2.1 presents a simulation of the level function $\theta(S(t), \beta(t))$ specified in the TDRS Vasicek model. In fact, the model has the capacity to include some interesting special cases, such as the time-homogeneous regime-switching models and permanent regime-switching models (Figure 2.2).

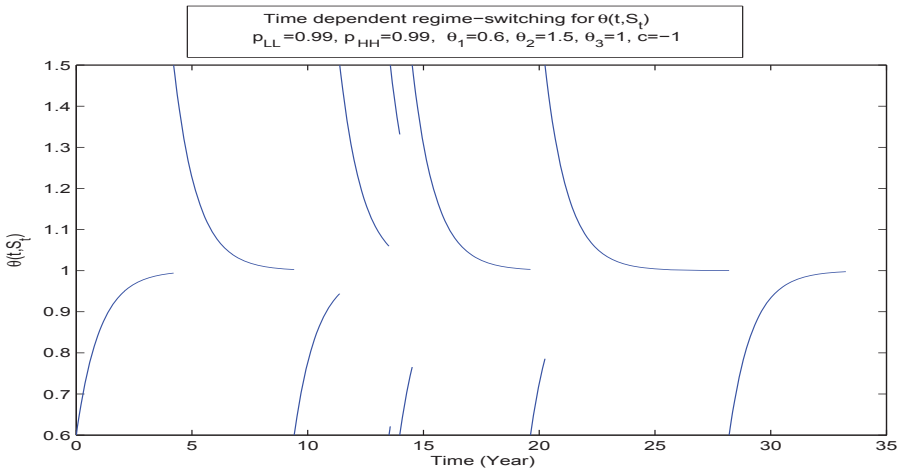


Figure 2.1: Simulated process of the mean-reverting level, where $\theta(S(t), \beta(t))$ is a piecewise exponential function of time

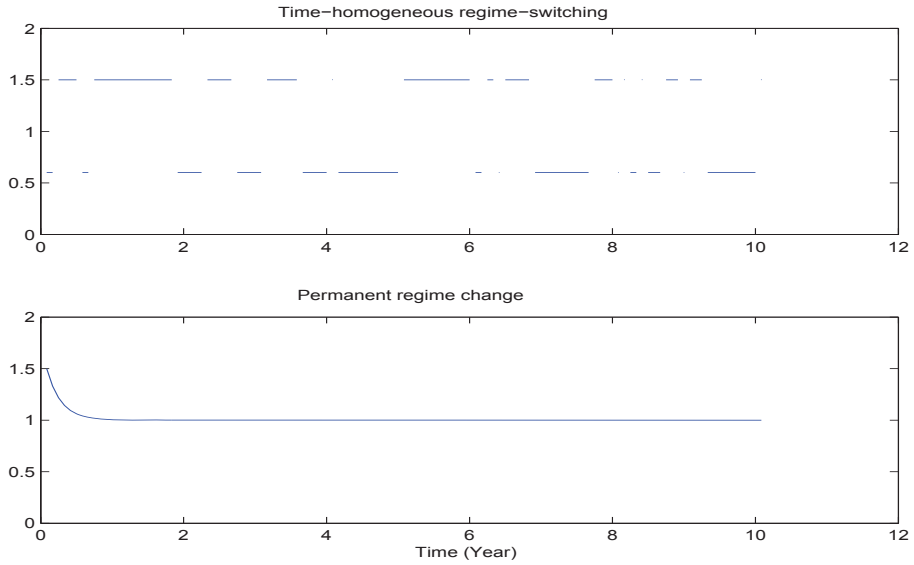


Figure 2.2: Level function for special cases of TDRS Vasicek models: the upper graph corresponds to the case $c = 0$; the lower graph corresponds to the case $p_{LL} = p_{HH} = 1$

Figure 2.3 provides weekly simulations of a short-rate process using the TDRS Vasicek model with two regimes.

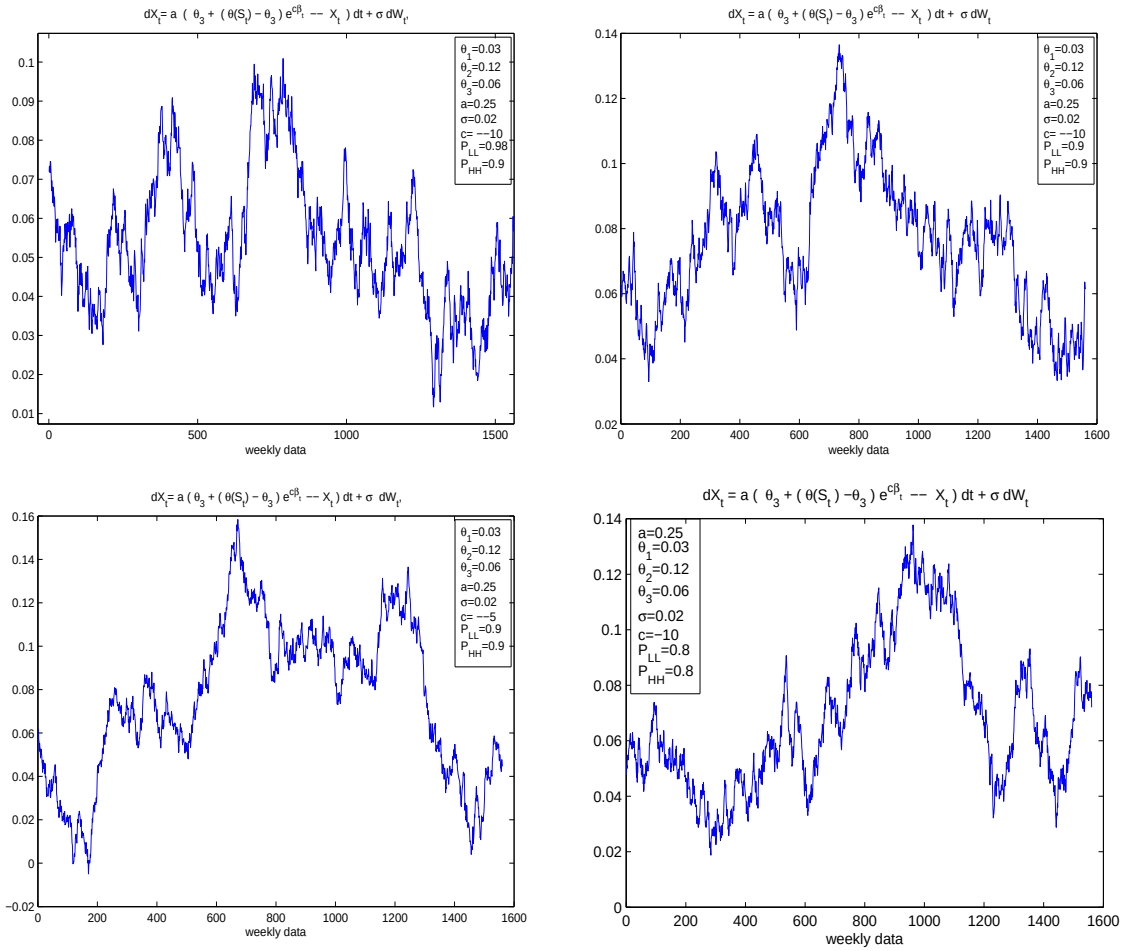


Figure 2.3: Simulated weekly data for 30 years

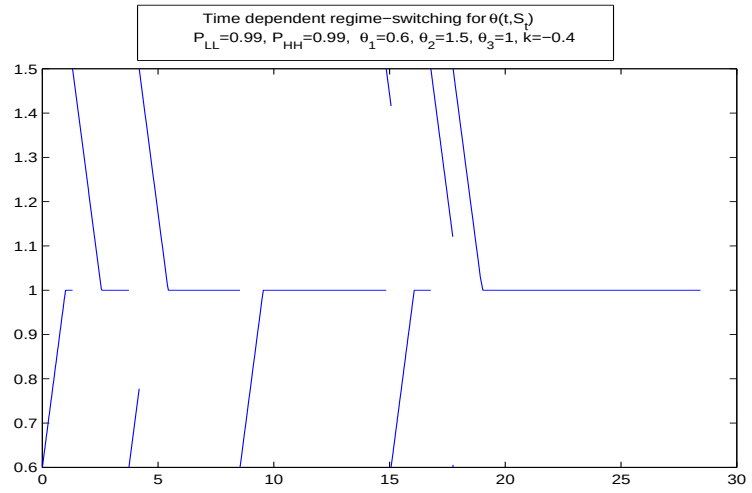
Moreover, in the TDRS General Vasicek model (2.4), one may specify a level function different from an exponential form, such as a linear form

$$\theta(S(t),\beta(t)) = \mu(S(t)) + k \cdot sgn(\mu(S(t)) - \theta_3) \cdot \min\{\beta(t), |\frac{\mu(S(t)) - \theta_3}{k}|\},$$

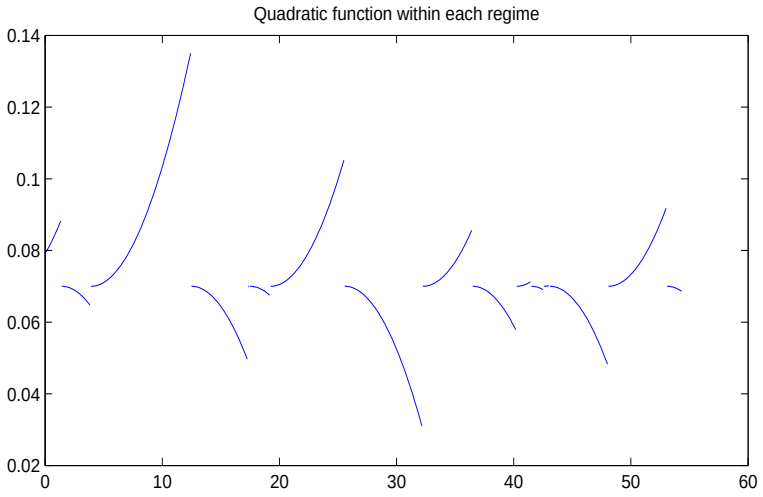
or a quadratic form

$$\theta(S(t),\beta(t)) = \theta_3 + sgn(\mu(S(t)) - \theta_3)c^2\beta(t)^2.$$

We present samples generated from these processes in Figure 2.4.



(a)



(b)

Figure 2.4: Different forms of level function: the upper graph corresponds to a linear form; the lower graph corresponds to a quadratic form

2.3 Methodology for Parameter Estimation

In Mao and Yuan (2006), the authors discuss some probabilistic properties of SDEs with Markovian switching, such as existence, uniqueness, methods of numerical approximations, boundedness, and stability. To the best of our knowledge, little work has been done on statistical inference on continuous-time SDEs with Markovian switching when the SDEs are time dependent. We have not found any mention about model (2.1) in the literature. Certainly, there are many interesting issues regarding this model, including both its probabilistic and statistical aspects. We initiate here the first effort to explore this new class of models, and we hope further work will be done in the future.

In general, there is no closed-form transition density function for (2.1). In fact, if we sample the continuous process at fixed discrete frequencies, we are facing random variables that involve integrals of the hidden continuous-time Markov chain and have no explicit analytic representation. One may approximate the likelihood function by using a discretization method. For example, we can approximate a regime-switching SDE by using the Euler discretization (Mao and Yuan, 2006) and then calculate the exact likelihood function of the discretized process as an approximation to the true likelihood function. As a result, we can use the discretized version of (2.1) as a possible way to estimate parameters.

In the context of non-regime-switching processes, numerous authors have adopted such an approach. Below we provide only some examples:

- Florens-Zmirou (1989) studies the following model

$$dX(t) = b(X(t), \theta)dt + \sigma dW(t),$$

  where θ and σ are unknown parameters. The author has proved that the MLE of θ and σ based on a discrete approximating scheme is consistent when the sampling interval shrinks to zero.

- Chan et al. (1992) applies the Generalized Method of Moments (GMM) to a discretized version of a continuous-time SDE.

- Stanton (1997) studies the following stationary univariate diffusion process:

$$dX(t) = \mu(X(t))dt + \sigma(X(t))dW(t).$$

  In the paper, the author estimates the functions $\mu(\cdot)$ and $\sigma(\cdot)$ nonparametrically, using a kernel method. According to the author, "as long as we sample the data monthly or better, the errors introduced by using approximations rather than the true drift and diffusion are extremely small, especially when compared with the likely magnitude of estimation error".

- McLeish and Kolkiewicz (1997) study high order approximations to diffusion processes and propose methods for estimating parameters and assessing goodness of fit.

- Driffill et al. (2002) discretize several regime-switching CIR processes and evaluate the predictive power of those models in terms of bond pricing.

- Fan et al. (2003) estimate the following time-dependent univariate diffusion model:

$$dX(t) = \{\alpha_0(t) + \alpha_1(t)X(t)\}dt + \beta_0(t)X(t)^{\beta_1(t)}dW(t). \tag{2.6}$$

  Their estimation method is based on the discretized version:

$$Y_{t_i} \approx \{\alpha_0(t_i) + \alpha_1(t_i)X_{t_i}\}\Delta_i + \beta_0(t_i)X_{t_i}^{\beta_1(t_i)}\sqrt{\Delta_i}\epsilon_{t_i}, \quad i = 1, \cdots, n. \tag{2.7}$$

Some of these approaches have been extended to the case of diffusion processes with regime-switching. For example, Choi (2009) investigates the time-homogeneous regime-switching SDE,

$$dX(t) = \mu(X(t), S(t); \theta)dt + \sigma(X(t), S(t); \theta)dW(t),$$

where $\mu(X(t), S(t); \theta) = \alpha_{-1S(t)} X(t)^{-1} + \alpha_{0S(t)} + \alpha_{1S(t)} X(t) + \alpha_{2S(t)} X(t)^2 + \alpha_{3S(t)} X(t)^3$ and $\sigma(X(t), S(t); \theta) = \beta_{S(t)} X(t)^{\rho_{S(t)}}$. The regime index $S(t)$ follows a continuous-time Markov chain with two states. The author uses a data set (weekly observed T-bill rates from 1971 to 2008) and approximates the true transition density function within each regime by the method proposed in Aït-Sahalia (2002). When estimating the parameter set θ, the author assumes that $\{S(t)\}$ follows a discrete-time Markov chain at the data observation times. This assumption is based on the observation that for small enough Δ, at most one regime shift can occur every seventh day.

Since the likelihood function based on the continuous-time model (2.1) is generally unknown, we are unable to find the maximum likelihood estimator for parameters in the drift and volatility coefficients. Therefore, we describe now a possible discretization of the model (2.1) and propose a parameter estimation method based on the discretized model. We assume that $f_\theta(X(t), S(t), \beta(t))$ is continuous as a function of $\beta(t)$. An Euler approximation to (2.1) is given by

$$X_{n+1} - X_n = f(X_n, S_n, \beta_n \Delta; \theta)\Delta + g(X_n, S_n, \beta_n \Delta; \theta)\sqrt{\Delta} Z_n, \quad n = 0, 1, \cdots \quad (2.8)$$

where $Z_n \sim N(0, 1)$, $X_n = X(n\Delta)$, $S_n = S(n\Delta)$ is a discrete-time Markov chain with N states and θ is a vector of unknown parameters; $\beta_n \triangleq n - \tau_n$, where τ_n is the last transition time of the discrete-time Markov chain $\{S_n\}$. Therefore, β_n stands for the number of intervals that $\{S_n\}$ has spent in the current regime. The fact that we calculate the duration β_n based on discrete observations should not impact significantly the estimation method we are proposing, because the error is at most a magnitude of Δ, and the function $f(\cdot)$ for each $S(t)$ is a continuous function of $\beta(t)$.

Let us define a two-dimensional Markov chain:

$$\alpha_n^0 = \begin{pmatrix} S_{n-1} \\ \beta_{n-1} \end{pmatrix}, \quad n = 1, 2, \cdots . \quad (2.9)$$

We are interested in filtering the process from observations of the model (2.1). We will convert $\{\alpha_n^0\}$ into a one-dimensional Markov chain so that we can use the filtering method proposed in Hamilton (1989). For this, we first construct a new Markov chain, denoted by $\{\alpha_n\}$, as follows:

$$\alpha_n = Nj + i, \text{ if } S_{n-1} = i, \beta_{n-1} = j. \quad 1 \leq i \leq N, j \geq 0, \tag{2.10}$$

where N is the number of regimes. The idea is to construct a one-to-one function between the one-dimensional subject α_n and the two-dimensional subject (S_{n-1}, β_{n-1}). The one-to-one correspondence can be checked by noticing that $S_{n-1} \equiv \alpha_n \ (\mod N)$ and $\beta_{n-1} = (\alpha_n - S_{n-1})/N$. Let $I_{\{\cdot\}}$ be the indicator function and $p_{ij} \triangleq P(S_n = j | S_{n-1} = i)$. For $\{\alpha_n\}$, we have the following properties:

$$P(\alpha_{n+1} = Nj_2 + i_2 | \alpha_n = Nj_1 + i_1) = p_{i_1 i_1} I_{\{i_2 = i_1, j_2 = j_1 + 1\}} + p_{i_1 i_2} I_{\{i_2 \neq i_1, j_2 = 0\}},$$

where $i_1, i_2 = 0, 1, \cdots, N$ and $j_1 \geq 0, j_2 \geq 0$. The idea is that if S_n does not switch regime from S_{n-1}, the only change will be $\beta_n = \beta_{n-1} + 1$; if S_n switches regime from S_{n-1}, we will have $\beta_n = 0$. The following describes the transition matrix for $\{\alpha_n\}$. If we denote by $\{p_{ij}\}$ the transition matrix for $\{S_n\}$ and define matrices A, B, O of size $N \times N$ as follows:

$$A = \begin{pmatrix} 0 & p_{12} & p_{13} & \cdots & p_{1N} \\ p_{21} & 0 & p_{23} & \cdots & p_{2N} \\ p_{31} & p_{32} & 0 & \cdots & p_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{N(N-1)} & 0 \end{pmatrix}$$

$$B = \begin{pmatrix} p_{11} & 0 & \ldots & 0 \\ 0 & p_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_{NN} \end{pmatrix}$$

$$O = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix},$$

then, the transition matrix of $\{\alpha_n\}$ can be represented as

$$Q = \begin{pmatrix} A & B & O & \cdots & \cdots & \cdots & \cdots & \cdots \\ A & O & B & O & \cdots & \cdots & \cdots & \cdots \\ A & O & O & B & O & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ A & O & \cdots & O & O & B & O & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \end{pmatrix} \tag{2.11}$$

Although the size of the above matrix appears large, most entries are zeros. Note that $\{\alpha_n\}$ has an infinite number of states, since β_n ranges over all nonnegative integer values. It is infeasible to apply Hamilton's algorithm (1989, 1990) to the time series model (2.8), since a computer can store only vectors of finite length. However, it is natural to approximate the infinite-state Markov chain with one with finite states. The idea is to choose a number "D" large enough so that $Prob(\beta_n > D) < \epsilon$ with a prescribed $\epsilon > 0$. We explain the idea in the case when $\{S_n\}$ has only two regimes. In this case we have the following inequality:

$$P(\textit{waiting time between regime switches } > D) \leq max(p_{LL}^D, p_{HH}^D),$$

where p_{LL} is the probability of staying in the "low" regime in one time step Δ and p_{HH} is the probability of staying in the "high" regime in one time step Δ. For example, if $p_{LL} = 0.6$ and $p_{HH} = 0.7$, then

$$P(\beta_n > 20) \leq max(0.6^{20}, 0.7^{20}) = 8 \times 10^{-4}.$$

We can now define an approximating Markov chain $\{\alpha_n^{(D)}\}$ with only a finite number of states. More specifically, we define first a two-dimensional truncated Markov chain

$$\alpha_n^{(D,0)} = \begin{pmatrix} S_{n-1} \\ \beta_{n-1}^{(D)} \end{pmatrix}, \tag{2.12}$$

where $\beta_{n-1}^{(D)} = min(\beta_{n-1}, D)$. The number of states for $\{\alpha_n^{(D,0)}\}$ is $2(D+1)$. Similar to (2.10), we can define a new one-dimensional Markov chain,

$$\alpha_n^{(D)} = Nj + i, \text{ if } S_{n-1} = i, \beta_{n-1} = j, \quad 1 \leq i \leq N, 0 \leq j \leq D. \tag{2.13}$$

Using the same matrix notation as before, we can find that the transition matrix of this process is of the form

$$Q^{(D)} = \begin{pmatrix} A & B & O & \cdots & O \\ A & O & B & \cdots & O \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ A & O & \cdots & O & B \\ A & O & O & \cdots & B \end{pmatrix}.$$

The truncated version of (2.8) is then

$$X_n = f_\theta(X_{n-1}, \alpha_n^{(D)})\Delta + g_\theta(X_{n-1}, \alpha_n^{(D)})\sqrt{\Delta}Z_n. \tag{2.14}$$

Remark 2.2. *Model (2.14) is a nonlinear regime-switching AR process. Therefore, there is a range of computational methods we can use to find the maximum likelihood estimate of parameters, including the EM algorithm, simulated likelihood, and direct likelihood.*

2.4 Method of Maximum Likelihood Estimation

In this section, we apply the EM algorithm to estimate parameters of the model (2.14). The EM algorithm is chosen over other computational methods such as the

simulated likelihood method and the direct likelihood method, due to the properties presented in Section 1.2.3 (direct likelihood and simulated likelihood turned out to be more computationally expensive in our pilot simulation studies).

To present the general steps for the EM algorithm, we need the following notation and definitions:

- D is the number of lags retained from the original Markov chain $\{\alpha_n\}$.
- M is the number of states of the Markov chain $\{\alpha_n^{(D)}\}$, i.e., $M = 2D + 2$.
- $I_n = \sigma\{X_i; i \leq n\}$ is the $\sigma-$field generated by $\{X_i; i \leq n\}$, representing the information contained in observed data up to time t_n.
- $f(x_n|\alpha_n^{(D)}, I_{n-1})$ is the conditional density function of X_n given $\alpha_n^{(D)}$ and I_{n-1}.
- $\eta_n = (f(x_n|\alpha_n^{(D)} = 1, I_{n-1}), \cdots, f(x_n|\alpha_n^{(D)} = M, I_{n-1}))'$.
- $1_M = (1, \cdots, 1, \cdots, 1)'$ is a vector of 1s with length M.
- $\hat{\xi}_{n|m} = (P(\alpha_n^{(D)} = 1|I_m), \cdots, P(\alpha_n^{(D)} = k|I_m), \cdots, P(\alpha_n^{(D)} = M|I_m))'$.
- $\lambda = (\theta, p_{LL}, p_{HH})'$ is the set of unknown parameters in model (2.14).

In the following, we present the EM algorithm described in Hamilton (1994) and later use it for our estimation purpose. Since the Markov chain $\{\alpha_n\}^{(D)}$ is not observable, we have to filter the process from observed data $\{x_n, n = 1, \cdots, T\}$. After the regimes are inferred from the data, we can calculate the log-likelihood function of λ and maximize it with the EM algorithm. The EM algorithm in our case is convenient because the first order conditions for maximizing the log-likelihood function can be explicitly solved. Here are the steps with some additional explanations:

Step 1: We start the filtering process with an input $\hat{\xi}_{1|1}$. We have two choices for $\hat{\xi}_{1|1}$. One is to set it equal to the stationary probability of $\{\alpha_n^{(D)}\}$ by the ergodic law of a Markov chain. The other is to treat $\hat{\xi}_{1|1} := \rho$ as a separate parameter vector to estimate. In this thesis it is set a parameter vector that will be inferred along with other parameters.

Step 2: The following equations will forecast the state variable $\alpha_{n+1}^{(D)}$ for the next step based on current information, I_n, and then update our inference on $\alpha_{n+1}^{(D)}$ when x_{n+1} is available:

$$\hat{\xi}_{n+1|n} = Q'^{(D)} \cdot \hat{\xi}_{n|n} \tag{2.15}$$

$$\hat{\xi}_{n+1|n+1} = \frac{\hat{\xi}_{n+1|n} \odot \eta_{n+1}}{1'(\hat{\xi}_{n+1|n} \odot \eta_{n+1})}, \tag{2.16}$$

where $Q'^{(D)}$ is the transpose of $Q^{(D)}$, and $\odot$ means the element-by-element product.

We would like to provide an additional explanation of this step. To see that (2.15) indeed holds, we may proceed as follows:

$$\begin{aligned}
& P(\alpha_{n+1}^{(D)} = i | I_n) \\
= & \sum_{j=1}^{M} P(\alpha_{n+1}^{(D)} = i, \alpha_n^{(D)} = j | I_n) \\
= & \sum_{j=1}^{M} P(\alpha_{n+1}^{(D)} = i | \alpha_n^{(D)} = j, I_n) P(\alpha_n^{(D)} = j | I_n) \\
= & \sum_{j=1}^{M} P(\alpha_{n+1}^{(D)} = i | \alpha_n^{(D)} = j) P(\alpha_n^{(D)} = j | I_n),
\end{aligned}$$

where the last equality holds because the conditional distribution of $\alpha_{n+1}^{(D)}$ given $\alpha_n^{(D)}$ is independent of I_n.

To see that (2.16) indeed holds,

$$\begin{aligned} & P(\alpha_{n+1}^{(D)} = i | I_{n+1}) \\ = \; & P(\alpha_{n+1}^{(D)} = i | I_n, x_{n+1}) \\ = \; & \frac{P(\alpha_{n+1}^{(D)} = i, x_{n+1} | I_n)}{P(x_{n+1} | I_n)} \\ = \; & \frac{P(x_{n+1} | \alpha_{n+1}^{(D)} = i, I_n) P(\alpha_{n+1}^{(D)} = i | I_n)}{P(x_{n+1} | I_n)} \\ = \; & \frac{P(x_{n+1} | \alpha_{n+1}^{(D)} = i, I_n) P(\alpha_{n+1}^{(D)} = i | I_n)}{\sum_{j=1}^{M} P(x_{n+1} | \alpha_{n+1}^{(D)} = j, I_n) P(\alpha_{n+1}^{(D)} = j | I_n)}. \end{aligned}$$

Step 3: Smoothed inference on the Markov chain $\{\alpha_n^{(D)}\}$ is obtained based on the whole observed trajectory $\{X_1, \cdots, X_T\}$, where T corresponds to the time for the last observation. In other words, we find the quantity $\hat{\xi}_{n|T} = (P(\alpha_n^{(D)} = 1|I_T), \cdots, P(\alpha_n^{(D)} = k|I_T), \cdots, P(\alpha_n^{(D)} = M|I_T))'$. The difference between this step and Step 2 is that in Step 2 we have only $\hat{\xi}_{n|n} = (P(\alpha_n^{(D)} = 1|I_n), \cdots, P(\alpha_n^{(D)} = k|I_n), \cdots, P(\alpha_n^{(D)} = M|I_n))'$, based on the observed historical trajectory up to time n. Now the algorithm starts from the very last value $\hat{\xi}_{T|T}$, which has been obtained at the end of Step 2, and then we iterate backwards to obtain $\hat{\xi}_{1|T}$. The algorithm presented here was developed by Kim (1993) and requires that the likelihood function of x_n depends on the latent Markov chain only through its current value $\alpha_n^{(D)}$. The procedure is based on the equation

$$\hat{\xi}_{n|T} = Q^{(D)} \odot \{\hat{\xi}_{n|n} \cdot (\hat{\xi}_{n+1|T}(\div)\hat{\xi}_{n+1|n})'\} \cdot 1_M, \tag{2.17}$$

where $(\div)$ denotes the element-by-element division.

Step 4: Since the Markov chain $\alpha_n^{(D)}$ is unobservable, we have an incomplete data problem. We can find the maximum likelihood estimators by using an EM algorithm described below.

- E step: The conditional expectation of the log-likelihood function of complete data $(\mathcal{X}, \alpha^{(D)})$ given the observed process $\mathcal{X}$ is

$$Q^{(D)}(\lambda_{l+1}; \lambda_l, \mathcal{X}, \alpha^{(D)}) = \sum_{\mathcal{S}} \log p(\mathcal{X}, \alpha^{(D)}; \lambda_{l+1}) p(\alpha^{(D)}|\mathcal{X}; \lambda_l), \quad (2.18)$$

  where $\mathcal{X} = (x_1, x_2, \cdots, x_T)$, $\alpha^{(D)} = (\alpha_1^{(D)}, \alpha_2^{(D)}, \cdots, \alpha_T^{(D)})$, and $\mathcal{S}$ is the set of all possible paths for $\alpha^{(D)}$.

- M step: By applying first order conditions to equation (2.18), we can derive normal equations.

Remark 2.3. *A byproduct of Hamilton's filtering method is that we can calculate the conditional likelihood function by*

$$f(x_n|I_{n-1}; \lambda) = 1'(\hat{\xi}_{n|n-1} \odot \eta_n). \quad (2.19)$$

However, the EM algorithm described in Step 4 *is more robust and saves computational cost, as documented in Hamilton (1990).*

2.5 TDRS Vasicek Model with Two Regimes

In this section, we would like to illustrate the estimation procedure explained in Section 2.3 by applying it to the TDRS Vasicek model. This model will later be employed for both simulation and application purposes.

2.5.1 The Maximum Likelihood Estimation

In order to use the Hamilton filtering method, we modify the model in the following two steps. First we discretize it. As a result, we obtain a Markov chain that may depend on an infinite number of past values. In the second step, we truncate the past data to the most recent D values. Below we describe the two steps in detail.

Time Series Analogue of the TDRS Vasicek Model

As explained in Section 2.3, if the time step is short enough, it suffices to consider at most one jump in each interval. Therefore, we can obtain the following time-series analogue of the TDRS Vasicek model:

$$X_{n+1} = X_n(1 - a\Delta) + a[\theta_3 + (\mu(S_n) - \theta_3)e^{c\beta_n\Delta}]\Delta + \sigma\sqrt{\Delta}\epsilon_n, \tag{2.20}$$

where the parameters have been introduced in (2.4) and (2.5), and β_n has been introduced in (2.8). In addition, we will use the following symbols for the transition probabilities:

$$\begin{aligned} p_{LL} &= P(S_n = 1|S_{n-1} = 1). \\ p_{HH} &= P(S_n = 2|S_{n-1} = 2). \\ p_{LH} &= P(S_n = 2|S_{n-1} = 1). \\ p_{HL} &= P(S_n = 1|S_{n-1} = 2). \end{aligned}$$

Truncated Time Series Model

Since we truncate β_n by a finite number D, our truncated time series will depend on the past D lags of S_n only (Section 2.3). As a result of the truncation, we obtain the following approximating process to model (2.20):

$$\hat{X}_{n+1} = \hat{X}_n(1 - a\Delta) + a[\theta_3 + (\mu(S_n) - \theta_3)e^{c\beta_n^{(D)}\Delta}]\Delta + \sigma\sqrt{\Delta}\epsilon_n, \tag{2.21}$$

where $\beta_n^{(D)} = \min(\beta_n, D)$ as defined in (2.12). It is obvious that $\beta_n^{(D)}$ can be determined from the previous D lags $(S_n, S_{n-1}, \cdots, S_{n-D})$.

By defining the Markov chain $\{\alpha_n^{(D)}\}$ as in (2.13), we can rewrite (2.21) to be a time-homogeneous regime-switching AR model :

$$\hat{X}_n = \hat{X}_{n-1}(1 - a\Delta) + a\omega(\alpha_n^{(D)})\Delta + \sigma\sqrt{\Delta}\epsilon_n, \tag{2.22}$$

where $\omega(\alpha_n^{(D)}) = \theta_3 + (\mu(S_{n-1}) - \theta_3)e^{c\beta_{n-1}^{(D)}\Delta}$.

In the two-regime case, we can write explicitly the transition probability matrix for $\alpha_n^{(D)}$:

$$Q^{(D)} = \begin{pmatrix} 0 & p_{LH} & p_{LL} & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ p_{HL} & 0 & 0 & p_{HH} & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & p_{LH} & 0 & 0 & p_{LL} & 0 & \cdots & \cdots & \cdots & 0 \\ p_{HL} & 0 & 0 & 0 & 0 & p_{HH} & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & p_{LH} & 0 & 0 & \cdots & 0 & 0 & 0 & p_{LL} & 0 \\ p_{HL} & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & p_{HH} \\ 0 & p_{LH} & 0 & 0 & \cdots & 0 & \cdots & \cdots & p_{LL} & 0 \\ p_{HL} & 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & p_{HH} \end{pmatrix}. \quad (2.23)$$

Consistency of the MLE

Under mild technical conditions, we can prove that there exists a stationary solution to the model (2.22) and that the maximum likelihood estimator for the parameters in (2.22) is also strongly consistent.

Theorem 2.1. *Assume that* $|1-a\triangle| < 1$ *and the transition probabilities* $p_{LL}, p_{HH} \in (0,1)$. *Then there exists a stationary solution to* (2.22). *Moreover, the maximum likelihood estimator for the parameters in (2.22) is strongly consistent.*

To prove the above theorem, we use the results by Francq and Roussignol (1998). To do so we have to verify that certain technical conditions are satisfied. We present a proof of this in Appendix 2.7.

The EM Algorithm

Define $Z_n \triangleq (I_{n-1}, \alpha_n^{(D)})$, meaning that Z_n contains the information given by I_{n-1} and $\alpha_n^{(D)}$. We have the log-likelihood function:

$$\log P(x_n|z_n;\theta) = C_1 - C_2 A_n^2, \tag{2.24}$$

with

$$\begin{aligned}
&C_1 = -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2}\log\Delta, \quad C_2 = \frac{1}{2\sigma^2\Delta} \\
&A_n = x_n - x_{n-1}(1-a\Delta) - a\Delta(\theta_3(1-e^{c\beta(\alpha_n^{(D)})\Delta}) + \mu(S_{n-1})e^{c\beta(\alpha_n^{(D)})\Delta}) \\
&\mu(2i-1) = \theta_1, \quad \mu(2i) = \theta_2, \quad i = 1, 2, \cdots, D+1 \\
&\beta(2i-1) = \beta(2i) = i-1, \quad i = 1, 2, \cdots, D+1.
\end{aligned}$$

The first derivatives of the log-likelihood function with respect to the unknown parameters are

$$\begin{aligned}
\frac{\partial \log P(x_n|z_n;\theta)}{\partial \sigma} &= -\frac{1}{\sigma} + \frac{A_n^2}{\sigma^3\Delta} \\
\frac{\partial \log P(x_n|z_n;\theta)}{\partial a} &= -\frac{A_n}{\sigma^2}[x_{n-1} - (\theta_3(1-e^{c\beta(\alpha_n^{(D)})\Delta})) + \mu(\alpha_n^{(D)})e^{c\beta(\alpha_n^{(D)})\Delta}] \\
\frac{\partial \log P(x_n|z_n;\theta)}{\partial \theta_3} &= \frac{A_n}{\sigma^2}a(1-e^{c\beta(\alpha_n^{(D)})\Delta}) \\
\frac{\partial \log P(x_n|z_n;\theta)}{\partial c} &= \frac{A_n}{\sigma^2}ae^{c\beta(\alpha_n^{(D)})\Delta}\beta(\alpha_n^{(D)})\Delta(\mu(\alpha_n^{(D)}) - \theta_3) \\
\frac{\partial \log P(x_n|z_n;\theta)}{\partial \theta_2} &= \frac{A_n}{\sigma^2}ae^{c\beta(\alpha_n^{(D)})\Delta}I_{\{\alpha_n^{(D)} \equiv 0 (\mathrm{mod}\ 2)\}} \\
\frac{\partial \log P(x_n|z_n;\theta)}{\partial \theta_1} &= \frac{A_n}{\sigma^2}ae^{c\beta(\alpha_n^{(D)})\Delta}I_{\{\alpha_n^{(D)} \equiv 1 (\mathrm{mod}\ 2)\}}.
\end{aligned}$$

Following the method presented in Section 2.4, we find that the updated pa-

rameter vector $\lambda^{(l+1)}$ must satisfy the following equations:

$$
\begin{aligned}
p_{HH}^{(l+1)} &= \frac{\sum_{n=2}^{T}\sum_{j=1}^{D} P(\alpha_n^{(D)}=2j+2,\alpha_{n-1}^{(D)}=2j|I_T;\lambda^{(l)})}{\sum_{n=2}^{T}\sum_{j=1}^{D+1} P(\alpha_{n-1}^{(D)}=2j|I_T;\lambda^{(l)})} \\
&\quad + \frac{P(\alpha_n^{(D)}=2D+2,\alpha_{n-1}^{(D)}=2D+2|I_T;\lambda^{(l)})}{\sum_{n=2}^{T}\sum_{j=1}^{D+1} P(\alpha_{n-1}^{(D)}=2j|I_T;\lambda^{(l)})} \\
p_{LL}^{(l+1)} &= \frac{\sum_{n=2}^{T}\sum_{j=1}^{D} P(\alpha_n^{(D)}=2j+1,\alpha_{n-1}=2j-1|I_T;\lambda^{(l)})}{\sum_{n=2}^{T}\sum_{j=1}^{D+1} P(\alpha_{n-1}^{(D)}=2j-1|I_T;\lambda^{(l)})} \\
&\quad + \frac{P(\alpha_n^{(D)}=2D+1,\alpha_{n-1}^{(D)}=2D+1|I_T;\lambda^{(l)})}{\sum_{n=2}^{T}\sum_{j=1}^{D+1} P(\alpha_{n-1}^{(D)}=2j-1|I_T;\lambda^{(l)})} \\
\rho_{i_1}^{(l+1)} &= P(\alpha_1^{(D)}=i_1|I_T;\lambda^{(l)}), i_1=0,\cdots,M
\end{aligned}
$$

$$
\sum_{n=2}^{T}\sum_{\alpha_n^{(D)}=0}^{M} \frac{\partial \log P(x_n|z_n;\theta)}{\partial\theta} \mid_{\theta=\theta^{(l+1)}} P(\alpha_n^{(D)}|I_T;\lambda^{(l)}) = 0,
$$

where $\lambda=(\theta,p_{LL},p_{HH},\rho)$, $\theta=(a,\theta_1,\theta_2,\theta_3,\sigma,c)$, $M=2D+2$, and $I_T=\{x_n\}_{1\le n\le T}$.

2.5.2 Estimation Results for Simulated and Real Data

Simulated Data

We simulate a sample path of 1440 monthly observations from model (2.5) with a piecewise exponential level function $\theta(S(t),\beta(t))$. The parameters we employ here are exactly the ones estimated from real monthly T bill data when applying the TDRS Vasicek model with two regimes (Table 2.3). We discard the first 720 observations so that the simulated path is approximately stationary. Therefore, 720 monthly observations are left. Then we estimate the marginal density function of the process by using the Matlab function *ksdensity*, which is based on a normal kernel function, with a window parameter ('width') that is a function of the number of observations. Figure 2.5 shows the estimated marginal density function.

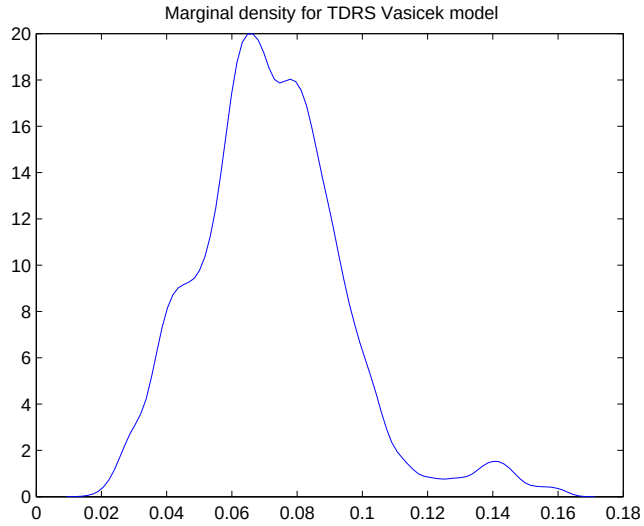


Figure 2.5: Estimated marginal density function from simulated data

One feature of the estimated marginal density function is that it is multi-modal, which can be attributed to the regime-switching feature and time-dependence in the model.

Now we present the estimation results based on the EM algorithm. The true parameter values and estimation values are presented in Tables 2.1 and 2.2, where we calculate the averages and standard errors of the maximum likelihood estimates from simulated samples, and D is the number of lags retained for the regime-switching process (Markov chain). It can be seen that most parameters can be estimated with relative errors less than 20%, where the relative error is calculated as the sample standard errror divided by true parameter value. Based on our simulation results, it seems that the choice of different values of D does not have significant impact on the estimated values. We also tried the EM algorithm with different initial values, and found that the EM algorithm is quite robust to the choice of initial parameters.

Table 2.1: EM algorithm based on 50 simulations of 60 years' monthly data, D=4

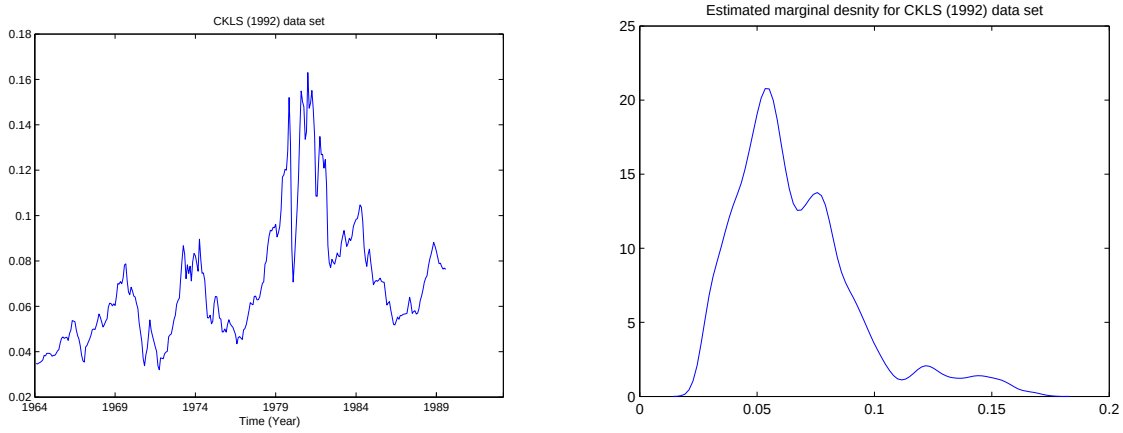
	True Value	Average Estimated Value	Sample Std. Err.
a	10	10.1847	0.8080
θ_1	0.6	0.5996	0.0337
θ_2	1.5	1.5186	0.2255
θ_3	1	1.0389	0.5201
p_{LL}	0.6	0.6221	0.0890
p_{HH}	0.6	0.6022	0.1185
σ	2	1.9493	0.1066
c	-5	-6.7109	5.2100

Table 2.2: EM algorithm based on 60 simulations of 60 years' monthly data, D=50

	True Value	Average Estimated value	Sample Std. Err.
a	10	10.5929	0.4658
θ_1	0.6	0.4764	0.2811
θ_2	1.5	1.5647	0.2816
θ_3	1	1.0064	0.1717
p_{LL}	0.9	0.8594	0.0903
p_{HH}	0.9	0.8525	0.1126
σ	2	1.8864	0.0858
c	-5	-4.7765	5.7991

Application to Interest Rates

The data set we investigate is monthly observed three-month US treasury bill data studied in the paper CKLS (1992). It consists of 307 monthly observations, from June 1964 to December 1989. In this section, we first describe some qualitative features of our data set from the time-series plot and estimated marginal density function. Then we fit the interest rate data with the TDRS Vasicek model (2.5) and estimate the parameters. Figure 2.6 (a) shows the plot of the data set, and Figure 2.6 (b) shows the estimated marginal density function for monthly T-bill data. From these figures, we can observe a few characteristics of the data set:



(a) (b)

Figure 2.6: CKLS (1992) data set

- There is a peak in the early 1980s, which is quite dramatic and abrupt. It seems natural to consider a regime-switching model.
- In the long run, it seems that the short rate fluctuates around a certain level. The trends can be described as increasing, then fluctuating around a certain high level, then decreasing.

- The graph of estimated marginal density function of interest rates suggests that the density is multi-modal. Since the estimate from simulated data based on the TDRS Vasicek model (Figure 2.5) can capture this feature, the TDRS Vasicek model seems to be an appropriate candidate for modeling interest rates.

To estimate the parameters for our TDRS Vasicek model (2.5), we set initial values $p_{LL} = 0.9, p_{HH} = 0.9$, $D = 20$ and use 10^{-4} as the tolerance level for convergence of the EM algorithm. The estimated parameters are listed in Table 2.3. The estimation results are robust to the initial values of parameters.

Table 2.3: Estimated parameters for CKLS (1992) data set

a	θ_1	θ_2	θ_3	p_{LL}	p_{HH}	σ	c
0.20	-1.33	0.94	0.089	0.99	0.88	0.017	-3.72

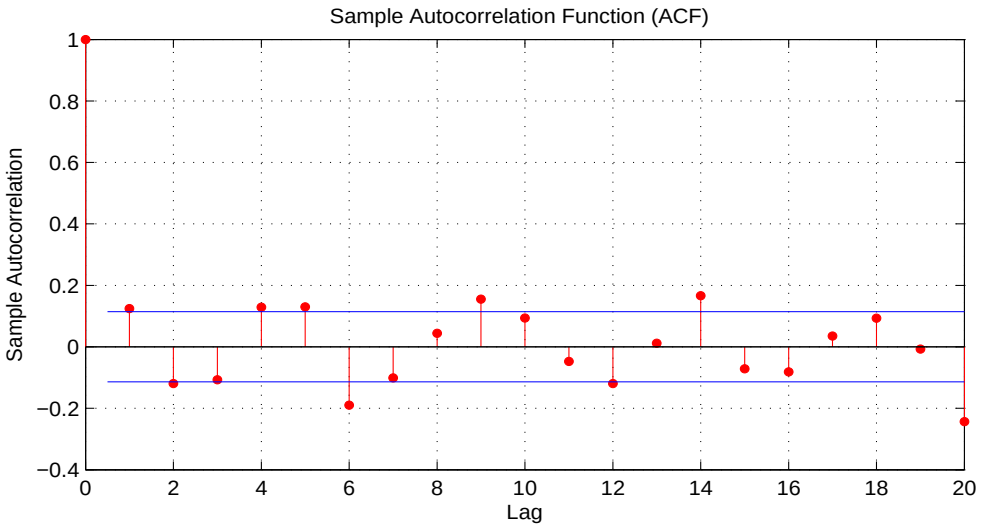
Using our estimated parameters from Table 2.3 and inferred regimes conditional on all observed data, we plot the autocorrelation function and normal QQ plot of the residuals from the truncated time series model (2.22) in Figure 2.7 (a) and (b). In Figure 2.7 (c), we plot the estimated mean-reversion level function $\theta(\hat{S}_t, \hat{\beta}_t)$, where $\hat{S}_t, \hat{\beta}_t$ are inferred conditionally on the entire history of observations. In addition, we plot the estimated conditional mean of $X(t)$ in Figure 2.7 (d). This can be obtained by first verifying, using Itô's formula, that

$$X(t) = e^{-at}X(0) + a\int_0^t e^{-a(t-u)}\theta(S(u), \beta(u))du + \sigma\int_0^t e^{-a(t-u)}dW(u)$$

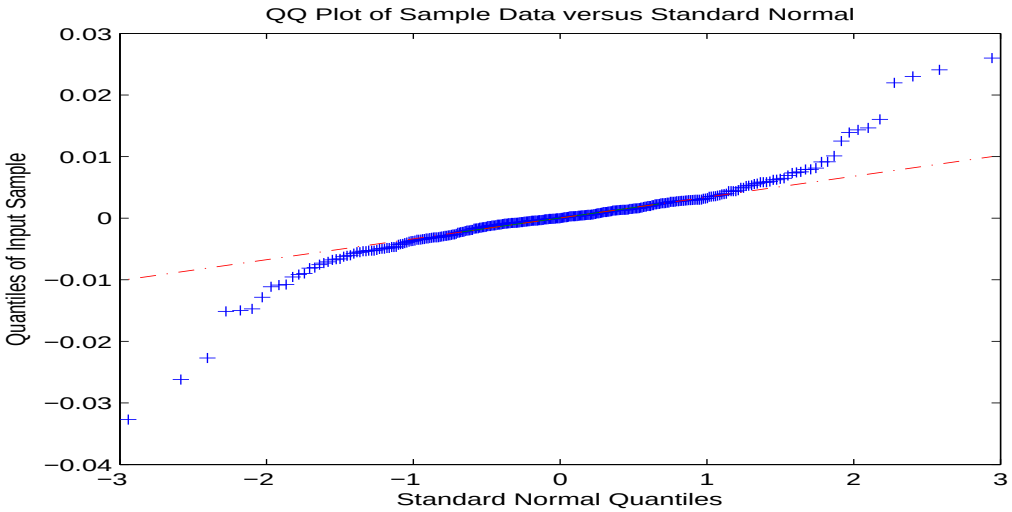
solves the TDRS Vasicek model. Then, by the independence of $\{W_u, 0 \leq u \leq t\}$ and $\{S(u), 0 \leq u \leq t\}$, we can find that

$$E[X(t)|X(0), S(u), 0 \leq u \leq t] = e^{-at}X(0) + a\int_0^t e^{-a(t-u)}\theta(S(u), \beta(u))du.$$

We call $E[X(t)|X(0), S(u), 0 \leq u \leq t]$ the conditional mean of $X(t)$, given a realization of $\{S(u), 0 \leq u \leq t\}$.

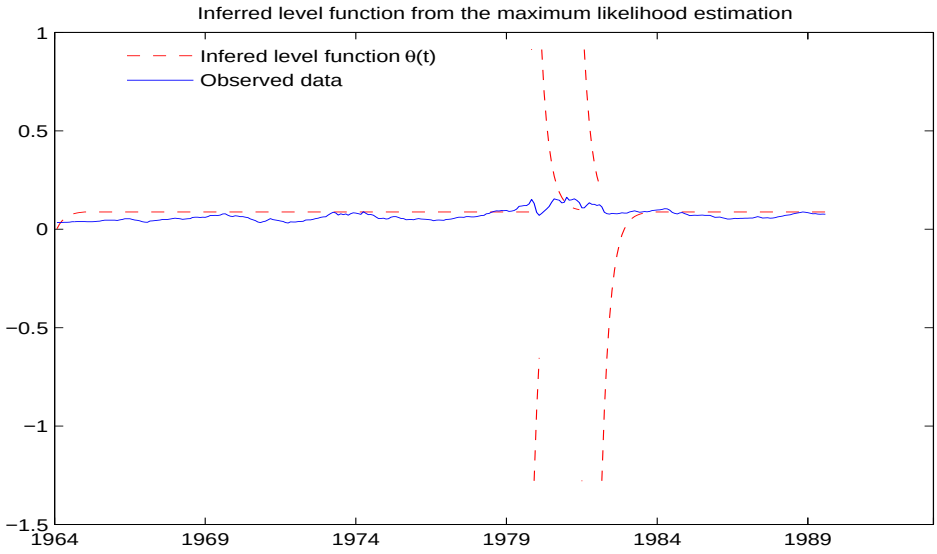


(a)

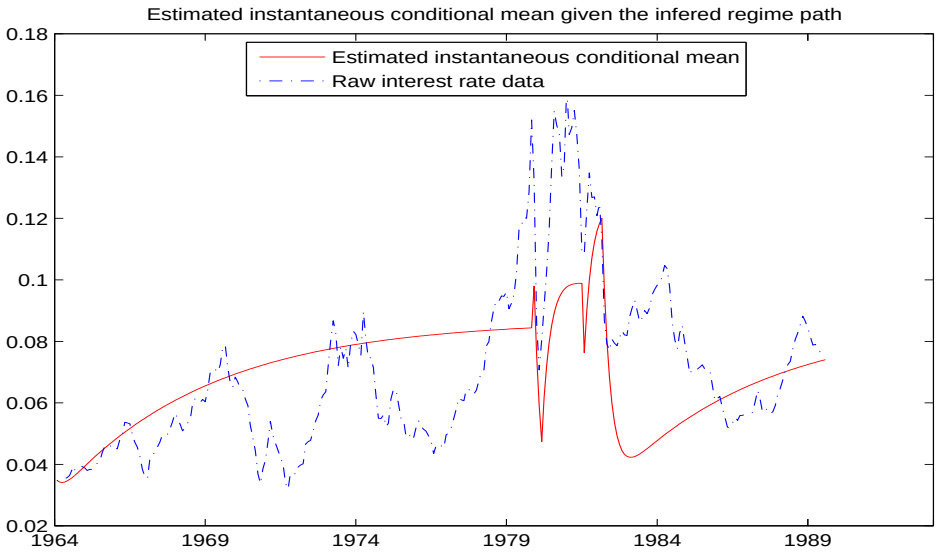


(b)

Figure 2.7: TDRS Vasicek model applied to T-bill data (CKLS 1992 data set)



(c)



(d)

Figure 2.7: TDRS Vasicek model applied to T-bill data (CKLS 1992 data set) (Continued)

From the estimation results, one can conclude the following:

- With the transition probabilities estimated in Table 2.3, we can calculate the stationary distribution of the hidden Markov chain. On average, the interest rate process spends 92% of its time in the lower regime and 8% of its time in the higher regime. Once the process comes into the lower regime, it takes eight years on average until it switches to the higher regime. However, it takes only eight months, or so, to shift to the lower regime after it gets into the higher regime.

- The graph suggests that there is time variation in the mean-reversion level function during the early 1980s, which corresponds to a monetary policy change. In October 1979, the United States Federal Reserve changed monetary policy from interest rate targeting to money supply targeting. The interest rate was very high in the late 1970s and can explain the abrupt regime shifts in our model. After October 1982, the monetary policy changed back to interest rate targeting and the regime shifted back to a lower regime that has continued until now.

- Figure 2.7 (a) suggests that the autocorrelation of residuals are not strong. In addition, from Figure 2.7 (b), the residuals seem to have fatter tails than normal distribution. If we allow more parameters, such as the mean-reverting speed and volatility coefficient, to depend on the regime changes, the data fit may be better. We consider some of these modifications in Chapter 5.

For this data set, we have also tried linear and quadratic forms of θ, as presented in Section 2.2. However, our algorithm does not give stable convergence results, suggesting that selection of the time-dependent component is a difficult problem.

2.6 Concluding Remarks

In this chapter, we have introduced a new class of time-dependent regime-switching models that allow the parameters to change with the time that has elapsed in the current regime. This dependency on elapsed time makes intuitive sense, because the calendar time does not have an obvious economic intuition behind it, while the elapsed time resets the time clock to zero once a new regime arrives. Another implication of the TDRS model is that it is stationary under some constraints, a situation explored in the next chapter. The stationarity of a model is a convenient feature as the asymptotic properties of traditional estimators are often easier to prove than non-stationary models.

The TDRS Vasicek model is a natural extension of the notable Vasicek model. The parametric specification of the level function $\theta(S(t), \beta(t))$ is not necessarily of exponential form, as illustrated in the application to the CKLS (1992) data set. We may also extend the TDRS Vasicek model to allow more parameters, such as the volatility coefficient σ, the mean reverting speed parameter a, and the parameter c in model (2.5), to depend on the hidden regime. The more flexible model would provide better goodness-of-fit of the data. These extensions can be explored in future research.

Our TDRS Vasicek model suggests significant time dependency in the early 1980s in the level of the diffusion process. In this model, the level function is of a specified parametric form and volatility is assumed to be constant. In Chapter 5, we consider a time-inhomogeneous model that leaves the level function unspecified and allows the volatility function to be estimated from the quadratic variation of the process. The empirical results in Chapter 5 confirm the findings of time-dependency of the level function in the early 1980s.

2.7 Appendix: Technical Proofs

Proof of Theorem 2.1

In (2.22), if we define functions $F(\cdot,\cdot)$ and $G(\cdot,\cdot)$ as follows:

$$\begin{aligned} F(x,i) &= (1-a\Delta)x + a[\theta_3 + (\mu(i)-\theta_3)e^{c\beta(i)\Delta}]\Delta \\ &= (1-a\Delta)x + a\omega(i)\Delta, \\ G(\epsilon_n,i) &= \sigma\sqrt{\Delta}\epsilon_n, \quad i = 1,2,\cdots,M, \end{aligned}$$

where $x \in \mathbb{R}$ and i is the value taken by the Markov chain $\{\alpha_n^{(D)}\}$, then it can be seen that (2.22) is a special case of the following general Markov-switching autoregressive time series model that has been studied in Francq and Roussignol (1998):

$$X_n = F(X_{n-1}, S_n, \lambda) + G(\eta_n, S_n, \lambda), \quad \forall n \geq 1, \tag{2.25}$$

where $\{\eta_n\}$ is a sequence of independent and identically distributed multivariate random vectors; λ is an unknown parameter belonging to an open subset Λ of $\mathbb{R}^d$; $\{S_n\}$ is a Markov chain independent of $\{\eta_n\}$ with finite state space $\mathbb{S} = \{1,2,\cdots,M\}$, and $F(\cdot)$ and $G(\cdot)$ are measurable functions. The authors give conditions for the existence of an ergodic stationary solution X_n to (2.25). They also prove the strong consistency of the maximum likelihood estimator.

We now present the technical conditions and show that these conditions are satisfied in our case. The authors assume that

A1 $\forall\lambda \in \Lambda$, the Markov chain $\{S_n\}$ is irreducible and aperiodic.

A2 $\forall\lambda \in \Lambda$, Equation (2.25) admits an ergodic stationary solution (and the observations are supposed to be generated by an ergodic stationary solution of (2.25)).

Suppose for each $i \in \mathbb{S}$ that $G(\eta_n, i, \theta)$ has density $f_{i,\theta}(\cdot)$, and $\underline{f}_\lambda(x,y) = \min_{1\leq i\leq M} f_{i,\lambda}(x - F(y,i,\lambda))$, $\overline{f}_\lambda(x,y) = \max_{1\leq i\leq M} f_{i,\lambda}(x - F(y,i,\lambda))$, $\underline{f}^*_\lambda(x) = \min_{1\leq i\leq M} f_{i,\lambda}(x - F(x_0,i,\lambda))$, and $\overline{f}^*_\lambda(x) = \max_{1\leq i\leq M} f_{i,\lambda}(x - F(x_0,i,\lambda))$. Then,

A3 The sets $\{(x,y) : f_{i,\lambda}(x-F(y,i,\lambda)) > 0\}$ and $\{(x,y) : f_{i,\lambda}(x-F(x_0,i,\lambda)) > 0\}$ do not depend on $i \in \mathbb{S}$ and x_0.

A4 $\forall\lambda \in \Lambda$ and $x_0 \in \mathbb{R}$, we have $E_{\lambda_0} sup_{||\lambda'-\lambda||<\delta} |log \underline{f}_{\lambda'}(X_n, X_{n-1})| < \infty$, $E_{\lambda_0} sup_{||\lambda'-\lambda||<\delta} |log \overline{f}_{\lambda'}(X_n, X_{n-1})| < \infty$, $E_{\lambda_0} sup_{||\lambda'-\lambda||<\delta} |log \underline{f}^*_{\lambda'}(X_n)| < \infty$, and $E_{\lambda_0} sup_{||\lambda'-\lambda||<\delta} |log \overline{f}^*_{\lambda'}(X_n)| < \infty$, for some $\delta > 0$.

A5 For all $(x,y) \in \mathbb{R} \times \mathbb{R}$, the functions $p_{i,j}(\cdot) = P(S_n = j | S_{n-1} = i)$ and $\lambda \mapsto f_{i,\lambda}(x - F(y,i,\lambda))$ are continuous over Λ.

A6 $\forall\lambda \in \Lambda$, if $g_{1,\lambda}(X_n|X_{n-1}, X_{t-2}, \cdots) = g_{1,\lambda_0}(X_n|X_{n-1}, X_{n-2}, \cdots)$ P_{θ_0}-a.s. then $\lambda = \lambda_0$.

A7 For all $i \in \mathbb{S}$, $E(||G(\eta_n, i)||) < \infty$, where $||\cdot||$ denotes the Euclidean norm.

A8 There exist $a_1, a_2, \cdots, a_M$ such that $\forall i \in \mathbb{S}$ and $\forall (x,y) \in \mathbb{R} \times \mathbb{R}$, $||F(x,i) - F(y,i)|| \leq a_i ||x-y||$ and the matrix

$$H := \begin{pmatrix} p_{1,1}a_1 & p_{2,1}a_1 & \cdots & p_{M,1}a_1 \\ p_{1,2}a_2 & p_{2,2}a_2 & \cdots & p_{M,2}a_2 \\ \vdots & \vdots & & \vdots \\ p_{1,M}a_M & p_{2,M}a_M & \cdots & p_{M,M}a_M \end{pmatrix}$$

has a spectral radius strictly less than 1.

By Theorem 1 in the paper, one needs to check only assumptions A1, A7 and A8 for the existence of an ergodic stationary solution. Note that $\{\alpha_n^{(D)}\}$ is an ergodic aperiodic Markov chain as long as $0 < p_{LL}, p_{HH} < 1$. A1 is satisfied. It follows from Schwartz inequality that $E[|\sigma\sqrt{\Delta}\epsilon_n|] \leq \sigma\sqrt{\Delta}\sqrt{E(\epsilon_n^2)} = \sigma\sqrt{\Delta} < +\infty$, and hence, A7 is satisfied. Moreover,

$$||F(x,i) - F(y,i)|| = |(1 - a\Delta)||x - y|.$$

Let

$$H = \begin{pmatrix} p_{1,1}(1-a\Delta) & p_{2,1}(1-a\Delta) & \cdots & p_{M,1}(1-a\Delta) \\ p_{1,2}(1-a\Delta) & p_{2,2}(1-a\Delta) & \cdots & p_{M,2}(1-a\Delta) \\ \vdots & \vdots & & \vdots \\ p_{1,M}(1-a\Delta) & p_{2,M}(1-a\Delta) & \cdots & p_{M,M}(1-a\Delta) \end{pmatrix}.$$

Then the spectral radius of H satisfies

$$\rho(H) \leq ||H||_\infty = \max_{1\leq i\leq M} \sum_{j=1}^{M} (1-a\Delta)|p_{ij}| = |1-a\Delta|,$$

where $||\cdot||$ is the induced ∞-norm. Therefore, if $|1-a\Delta| < 1$, then A8 is satisfied.

By Theorem 3 in the paper, one needs to check conditions A1–A6 for the consistency of MLE. We have shown that under the conditions $|1-a\Delta| < 1$ and $0 < p_{LL}, p_{HH} < 1$, A1, A7 and A8 are satisfied. Therefore, there exists an ergodic stationary solution to (2.22) and $A2$ is satisfied Note that $G(\epsilon_n, i) = \sigma\sqrt{\Delta}\epsilon_n$ has density function $f_{i,\lambda}(x) = \frac{1}{\sqrt{2\pi\Delta}\sigma} exp\{-\frac{x^2}{2\Delta\sigma^2}\}$. Obviously, $\{(x,y) : f_{i,\lambda}(x - F(y,i,\lambda)) > 0\} = \mathbb{R}^2$ and $\{x : f_{i,\lambda}\ (x - F(x_0,i,\lambda)) > 0\} = \mathbb{R}$. Therefore, A3 is satisfied. By definition,

$$\begin{aligned} \underline{f}_{-\lambda'}(X_n, X_{n-1}) &= \min_{1\leq i\leq M} f_{i,\lambda'}(X_n - F(X_{n-1}, i, \lambda')) \\ &= \min_{1\leq i\leq M} \frac{1}{\sqrt{2\pi\Delta}\sigma'} exp\{-\frac{X_n - F(X_{n-1}, i, \lambda')^2}{2\Delta\sigma'^2}\}. \end{aligned}$$

Provided that $E_{\theta_0}|X_n|^2 < \infty$, for each $\lambda \in \Lambda = (a > 0, 0 < \theta_1 < \theta_3 < \theta_2, 0 < p_{LL}, p_{LH}, p_{HH}, p_{HL} < 1, \sigma > 0)$, there exists $\delta > 0$ such that

$$E_{\theta_0} \sup_{||\lambda'-\lambda||<\delta} ||log f_{-\lambda'}(X_n, X_{n-1})|| < \infty.$$

Now,

$$X_n = (1-a\Delta)X_{n-1} + a\omega\Delta(\alpha_n^{(D)}) + \sigma\sqrt{\Delta}\epsilon_n.$$

By recursively applying the above equation we get

$$X_n = \sum_{i=0}^{\infty} (1 - a\Delta)^i [a\omega(\alpha_{n-i}^{(D)})\Delta + \sigma\sqrt{\Delta}\epsilon_{n-i}].$$

By the Cauchy criterion, it can be shown that $E[X_n^2] < \infty$ under the condition $|1 - a\Delta| < 1$. Therefore, if $|1 - a\Delta| < 1$, then A4 is satisfied. A5 is obviously satisfied. Let us check the last condition A6, which deals with the identifiability of the parameters. Note that the set of parameters $(\theta_1, \theta_2, \theta_3, c)$ can be uniquely determined by the values of $\{\omega(\alpha_n^{(D)})\}$ as long as $D \geq 2$ (the number of equations should be more than the number of parameters). By routine and tedious application of the approach presented in the example given in Francq and Rossignol (1998), it can be verified that A6 is satisfied. □

Chapter 3

Theoretical Properties of the Proposed Model

3.1 Introduction

Whenever a new model is proposed, it is helpful to investigate its analytical properties, such as the existence, uniqueness, and stationarity. To serve such a purpose, we devote this chapter to proving certain important theoretical properties of the proposed TDRS and TDRS Vasicek models. The key assumption behind the results in this chapter is that the two sources of randomness, the hidden Markov chain and the Brownian motion, are independent of each other. To show the existence and uniqueness of a solution to the TDRS model, the idea is to simulate a trajectory of the Markov chain first and then show the existence of a solution to the SDE conditional on the path of the Markov chain. To show the stationarity of the TDRS Vasicek model, the idea is to decompose the solution of the model into a sum of two independent pieces, one involving only Brownian motion and the other involving only the hidden Markov chain. The stationarity of the TDRS model then

follows from the stationarity of each component of the solution. We consider the results presented in this chapter to be an extension of the techniques in Mao and Yuan (2006), and we document connections between our contributions and Mao and Yuan's ideas as we present them.

The layout of the rest of this chapter is as follows:

- Section 3.2 reviews relevant background knowledge and tools that will be used in our proofs, such as the weak convergence of probability measures, and the existence and uniqueness of solutions to SDEs and SDEs with Markovian switching.
- Section 3.3 presents theoretical properties of TDRS models, including existence and uniqueness.
- Section 3.4 proves the stationarity of the TDRS Vasicek model and presents results on the moment behavior of the model.
- Section 3.5 draws conclusion for this chapter.

3.2 Preliminaries

First we review definitions and results needed in the proofs of our results in Section 3.4.

Definition 3.1. *A* **Polish space** $(X, \mathcal{B})$ *is defined to be a complete separable metric space* X *equipped with its Borel* $\sigma-$*algebra* $\mathcal{B}$*, i.e., the smallest* $\sigma-$*algebra generated by the sets that are open in its metric topology.*

Let $(X, \mathcal{B})$ be a Polish space. Then it is known that the set $\Delta(X, \mathcal{B})$ of all Borel probability measures on the measurable space $(X, \mathcal{B})$ is also a Polish space,

provided that it is given the topology of weak convergence of probability measures. This topology corresponds to Prohorov metric ρ, which defines a distance between any pair of probability measures $\mu, \nu \in \Delta(X, \mathcal{B})$ in the following way:

$$d_{Pro}(\mu, \nu) := \inf_{\epsilon}\{\epsilon > 0 | \forall E \in \mathcal{B}(X) : \mu(E) \leq \nu(N_\epsilon(E)) + \epsilon \text{ and } \nu(E) \leq \mu(N_\epsilon(E)) + \epsilon\},$$

where $N_\epsilon(E)$ denotes the set of all points in X within a distance $\epsilon > 0$ of points in E. This topology of weak convergence of probability measures derives its name from the property that a sequence of measures $(\mu_n)_{n=1}^{\infty}$ in $\Delta(X, \mathcal{B})$ converges to the limit $\mu \in \Delta(X, \mathcal{B})$ if and only if for every bounded continuous function $f : X \to \mathbb{R}$ the expected value $\int_X f(x)\mu_n(dx)$ of f with respect to the probability measure μ_n converges to the expected value $\int_X f(x)\mu(dx)$ of f with respect to the probability measure μ [Barberá et al., 2003].

Definition 3.2. *Let $(X, \mathcal{B})$ be a Polish space and $P_1, P_2 \in \Delta(X, \mathcal{B})$. Consider a complete metric d_X on X that is bounded by one. The* **bounded Lipschitz metric** *is defined by $d_{BL}(P_1, P_2) := \sup_f |E_{P_1} f - E_{P_2} f|$, where the supremum is taken over all functions f satisfying the Lipschitz condition $|f(x) - f(y)| \leq d_X(x, y)$.*

Theorem 3.1. *[Steinwart et al., 2008]. Let $(X, \mathcal{B})$ be a Polish space. If $P_1, P_2 \in \Delta(X, \mathcal{B})$, then*

$$d^2_{Pro}(P_1, P_2) \leq d_{BL}(P_1, P_2) \leq 2d_{Pro}(P_1, P_2).$$

It follows from the above theorem that the Prohorov metric and the bounded Lipschitz metric are topologically equivalent. Therefore, a sequence of probability measures converges under the Prohorov metric if and only if it converges under the bounded Lipschitz metric.

Definition 3.3. *A Markov process $\{X(t)\}$ is said to be* **homogeneous**, *if $P(X(t+s) \in B | X(s) \in A) = P(X(t) \in B | X(0) \in A)$, $\forall s, t \geq 0$.*

Example 3.1. *Let $\{X(t)\}$ be a strong solution to the following stochastic differential equation*

$$dX(t) = f(X(t))dt + g(X(t))dW(t), \tag{3.1}$$

where $f(\cdot)$, and $g(\cdot)$ satisfy the conditions to ensure the existence and uniqueness of a strong solution. Then $\{X(t)\}$ is a homogeneous Markov process.

In Sections 3.2.1 and 3.2.2, we review results on the existence and uniqueness of solutions to SDEs and SDEs with Markovian switching.

3.2.1 Stochastic Differential Equations

The results in this section are from Mao and Yuan (2006). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space with a filtration $\{\mathcal{F}(t)\}_{t\geq 0}$ satisfying the usual conditions. Let $W(t) = (W_1(t), \cdots, W_m(t))', \quad t \geq 0$ be an m-dimensional $\mathcal{F}(t)$-adapted Brownian motion defined on the space. For $0 \leq t_0 < T < \infty$, let $X(0) \in L^2_{\mathcal{F}_{t_0}}(\Omega; \mathbb{R}^n)$, i.e., $X(0)$ is an $\mathcal{F}_{t_0}$-measurable $\mathbb{R}^n$-valued random variable such that $E|X(0)|^2 < \infty$, where $|X(0)|$ is the Euclidean norm of $X(0)$. Define $\mathcal{M}^2([t_0, T]; \mathbb{R}^n)$ to be the set of $\{X(t)\}$ such that $\mathbb{E}(\int_{t_0}^T |X(s)|^2 ds) < \infty$. Let $f : \mathbb{R}^n \times [t_0, T] \to \mathbb{R}^n$ and $g : \mathbb{R}^n \times [t_0, T] \to \mathbb{R}^{n\times m}$ be Borel measurable. Consider the following n-dimensional stochastic differential equation of Itô type

$$dX(t) = f(X(t), t)dt + g(X(t), t)dW(t), \quad t_0 \leq t \leq T, \tag{3.2}$$

with the initial value $X(t_0) = X(0)$. By the definition of stochastic differential, this equation is equivalent to the following stochastic integral equation:

$$X(t) = X(0) + \int_{t_0}^t f(X(s), s)ds + \int_{t_0}^t g(X(s), s)dW(s) \quad \forall t \in [t_0, t]. \tag{3.3}$$

Definition 3.4. *An $\mathbb{R}^n$-valued stochastic process $\{X(t)\}_{t_0\leq t\leq T}$ is called a solution of (3.2) if it has the following properties:*

(1) *The process* $\{X(t)\}$ *is continuous and* $\mathcal{F}(t)$*-adapted;*

(2) $\{f(X(t),t)\} \in \mathcal{L}^1([t_0,T];\mathbb{R}^n)$ *and* $\{g(X(t),t)\} \in \mathcal{L}^2([t_0,T];\mathbb{R}^{n\times m})$*, where* $\mathcal{L}^P([a,b];\mathbb{R}^n)$ *denotes the family of* $\mathbb{R}^n$*-valued* $\mathcal{F}(t)$*-adapted processes* $\{f(t)\}_{a\le t\le b}$ *such that* $\int_a^b |f(t)|^p dt < \infty$ *a.s.;*

(3) *Equation (3.3) holds with probability* 1*.*

A solution $\{X(t)\}$ is said to be *unique* if any other solution $\{\bar{X}(t)\}$ is indistinguishable from $\{X(t)\}$; that is,

$$\mathbb{P}\{X(t) = \bar{X}(t) \text{ for all } t_0 \le t \le T\} = 1.$$

Theorem 3.2. *Assume that there exists a positive constant* K *such that*
(i) (Lipschitz condition) for all $x, y \in \mathbb{R}^n$ *and* $t \in [t_0, T]$*,*

$$|f(x,t) - f(y,t)|^2 \vee |g(x,t) - g(y,t)|^2 \le K|x-y|^2; \tag{3.4}$$

(ii)(Linear growth condition) for all $(x,t) \in \mathbb{R}^n \times [t_0,T]$*,*

$$|f(x,t)|^2 \vee |g(x,t)|^2 \le K(1+|x|^2). \tag{3.5}$$

There then exists a unique solution process $\{X(t)\}$ *to equation (3.2), and the solution belongs to* $\mathcal{M}^2([t_0,T];\mathbb{R}^n)$*.*

The following remarks are from Mao and Yuan (2006):

(a) The coefficients f and g can depend on ω in a general manner as long as they are adapted.

(b) Both initial time t_0 and final time T can be random variables provided they are stopping times.

(c) In the above results, we require the initial value $X(0)$ to be L^2, but in general, it is enough for $X(0)$ to be a random variable that is $\mathcal{F}_{t_0}$-measurable.

3.2.2 SDEs with Markovian Switching

The results in this section are from Mao and Yuan (2006). For the rest of this chapter, we assume that $\{S(t)\}$ is a right-continuous Markov chain on the probability space taking values in the finite state space $\mathbb{S} = \{1, 2, \cdots, N\}$ with generator $\Gamma = (\gamma_{ij})_{N\times N}$ given by

$$\mathbb{P}\{S(t+\Delta) = j | S(t) = i\} = \begin{cases} \gamma_{ij}\Delta + o(\Delta) & \text{if } i \neq j, \\ 1 + \gamma_{ii}\Delta + o(\Delta) & \text{if } i = j, \end{cases}$$

where $\Delta > 0$ and $\sum_{j\in\mathbb{S}} \gamma_{ij} = 0$ with $\gamma_{ij} \geq 0$ being the transition rate from i to j, $i \neq j$. We assume that the continuous-time Markov chain $\{S(t)\}$ is independent of the Brownian motion $\{W(t)\}$. Consider an SDE with Markovian switching of the form

$$dX(t) = f(X(t), t, S(t))dt + g(X(t), t, S(t))dW(t), \quad t_0 \leq t \leq T, \tag{3.6}$$

with initial value $X(t_0) = X(0) \in L^2_{\mathcal{F}_{t_0}}(\Omega; \mathbb{R}^n)$ and $S(t_0) = S_0$, where S_0 is an $\mathbb{S}$-valued $\mathcal{F}_{t_0}$-measurable random variable and

$$f : \mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{S} \to \mathbb{R}^n \text{ and } g : \mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{S} \to \mathbb{R}^{n\times m}$$

Definition 3.5. *An $\mathbb{R}^n$-valued stochastic process $\{X(t)\}_{t_0\leq t\leq T}$ is called a solution of equation (3.6) if it has the following properties:*

(1) *$\{X(t)\}_{t_0\leq t\leq T}$ is continuous and $\mathcal{F}(t)$-adapted.*

(2) *$\{f(X(t), t, S(t))\}_{t_0\leq t\leq T} \in \mathcal{L}^1([t_0, T]; \mathbb{R}^n)$; while $\{g(X(t), t, S(t))\}_{t_0\leq t\leq T} \in \mathcal{L}^2([t_0, T]; \mathbb{R}^n)$;*

(3) *for any $t \in [t_0, T]$, equation*

$$X(t) = X(0) + \int_{t_0}^{t} f(X(u), u, S(u))du + \int_{t_0}^{t} g(X(u), u, S(u))dW(u)$$

holds with probability 1.

Theorem 3.3. *Assume that there exists a positive constant K such that*
(i) (Lipschitz condition) for all $x, y \in \mathbb{R}^n$, $t \in [t_0, T]$ and $i \in \mathbb{S}$

$$|f(x,t,i) - f(y,t,i)|^2 \vee |g(x,t,i) - g(y,t,i)|^2 \leq K|x-y|^2; \tag{3.7}$$

(ii) (Linear growth condition) for all $(x,t,i) \in \mathbb{R}^n \times [t_0,T] \times \mathbb{S}$

$$|f(x,t,i)|^2 \vee |g(x,t,i)|^2 \leq K(1+|x|^2). \tag{3.8}$$

Then there exists a unique solution $X(t)$ to equation (3.6) and, moreover,

$$\mathbb{E}(\sup_{t_0 \leq t \leq T} |X(t)|^2) \leq (1 + 3\mathbb{E}|X(0)|^2)e^{3K(T-t_0)(T-t_0+4)}.$$

Thus, the solution belongs to $\mathcal{M}^2([t_0,T];\mathbb{R}^n)$.

3.3 Time-Dependent Regime-Switching Model

In this section, we investigate our general TDRS model (2.1) and present some of its theoretical properties, including the existence and uniqueness of a solution to model (2.1) and the limiting distribution of the unobserved process. The technical proofs for the results in this section are presented in Section 3.6.

For convenience, we remind the reader of the definition of our model (2.1):

$$\begin{aligned} dX(t) &= f(X(t),S(t),\beta(t))dt + g(X(t),S(t),\beta(t))dW(t), \quad t_0 \leq t \leq T \\ \beta(t) &: \quad \textit{the time that has elapsed in the current regime,} \end{aligned}$$

with initial values $X(t_0) = X(0) \in L^2_{\mathcal{F}_{t_0}}(\Omega;\mathbb{R}), \beta(t_0) = \beta_0 \in \mathbb{R}^+$, and $S(t_0) = S_0$; where S_0 and β_0 are $\mathcal{F}_{t_0}$-measurable random variables, and

$$f : \mathbb{R} \times \mathbb{S} \times \mathbb{R}_+ \to \mathbb{R} \text{ and } g : \mathbb{R} \times \mathbb{S} \times \mathbb{R}_+ \to \mathbb{R}.$$

Theorem 3.4. *Assume that there exists one positive constant K such that*
(i) (Lipschitz condition) for all $x, y \in \mathbb{R}$, $t \in [t_0, T]$ and $i \in \mathbb{S}$,

$$|f(x,i,\beta(t)) - f(y,i,\beta(t))|^2 \vee |g(x,i,\beta(t)) - g(y,i,\beta(t))|^2 \leq K|x-y|^2;$$

(ii) (Linear growth condition) for all $(x,i,t) \in \mathbb{R} \times S \times [t_0, T]$,

$$|f(x,i,\beta(t))|^2 \vee |g(x,i,\beta(t))|^2 \leq K(1+|x|^2).$$

Then there exists a unique solution $X(t)$ to equation (2.1), and

$$\mathbb{E}(\sup_{t_0 \leq t \leq T} |X(t)|^2) \leq (1 + 3\mathbb{E}|X(0)|^2)e^{3K(T-t_0)(T-t_0+4)}. \tag{3.9}$$

So the solution belongs to $\mathcal{M}^2([t_0, T]; \mathbb{R})$.

The idea behind the proof of the above theorem is to use the fact that $\{S(t)\}$ is a random constant between the jump times. Since the jump times are stopping times, we can find the solution to (2.1) between each pair of jump times and then "glue" them together over the entire interval $[t_0, T]$. The technique of the proof is similar to the one we used in proving the existence and uniqueness of a solution to (3.6). Let $\{\beta(t)\}$ be defined as in Section 2.1 and $\Pi = (\pi_i)$ be the stationary distribution of $\{S(t)\}$. For the rest of this section, we present some properties of the two-dimensional hidden process $(S(t), \beta(t))$. These properties are useful for the results in the next section as well. The following theorem states the Markov property and limiting distribution of the two-dimensional hidden process $(S(t), \beta(t))$ in model (2.1).

Theorem 3.5. *The process $\alpha(t) = (S(t), \beta(t))'$ is a two-dimensional homogeneous Markov process with the following limiting distribution: for each $x, h \in \mathbb{R}^+$ and $i \in \mathbb{S}$,*

$$\lim_{t\to\infty} P(\beta(t) \leq x, S(t) = i | S(u), \beta(u), u \in [0,h]) = \pi_i(1 - \exp(\gamma_{ii}x)).$$

.

Corollary 3.1. *Let $Y(t) = f(S(t), \beta(t))$, where $f : (\mathbb{S}, \mathcal{P}(\mathbb{S})) \otimes (\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is measurable and injective, and $\mathcal{P}(\mathbb{S})$ is the power set of the state space of $S(t)$. Then $\{Y(t)\}$ is a homogeneous Markov process.*

In the TDRS Vasicek model introduced in Section 2.2, the mean-reversion level function is of the following form:

$$\theta(S(t), \beta(t)) = \theta_3 + (\mu(S(t)) - \theta_3)e^{c\beta(t)}.$$

In this case, the range of possible values of $\theta(S(t) = 2, \beta(t))$ is disjoint with the range of $\theta(S(t) = 1, \beta(t))$. Moreover, it is easy to verify that $\theta(S(t), \beta(t))$ is an injective function from $\mathbb{S} \times \mathbb{R}_+$ to $[\theta_1, \theta_2]$. Therefore, $\theta(S(t), \beta(t))$ is a homogeneous Markov process by Corollary 3.1. In addition, we can find the density function of θ. If we denote the infinitesimal generator for the continuous-time Markov chain $\{S(t)\}$ by $\begin{pmatrix} -\gamma_{11} & \gamma_{11} \\ \gamma_{22} & -\gamma_{22} \end{pmatrix}$, then we can show that the limiting density function for $\theta(S(t), \beta(t))$:

$$p(y) = \begin{cases} \pi_1\gamma_{11}\left(\frac{\theta_1-\theta_3}{y-\theta_3}\right)^{\frac{\gamma_{11}}{c}} \frac{1}{c(y-\theta_3)}, & \theta_1 \le y < \theta_3 \\ -\pi_2\gamma_{22}\left(\frac{\theta_2-\theta_3}{y-\theta_3}\right)^{\frac{\gamma_{22}}{c}} \frac{1}{c(y-\theta_3)}, & \theta_3 < y \le \theta_2. \end{cases}$$

Thus if the process θ runs for a long time, then its marginal distribution has the density of the above form.

Now we consider a discretized version of $\alpha(t) = (S(t), \beta(t))'$, i.e., $\alpha_n^0 = (S_{n-1}, \beta_{n-1})'$ as defined in (2.8) and (2.9), where $\{S_n\}$ is a discrete-time Markov chain. Note that here $\beta_n \ge 0$ takes integer values only. It turns out that we have results similar to those in the continuous case.

Theorem 3.6. *$\{\alpha_n^0\}$ is a two-dimensional Markov chain with the following limiting distribution:*

$$\lim_{n\to\infty} P(S_n = i, \beta_n = j | S_0 = i_0, \beta_0 = j_0) = \pi_i P_{ii}^j (1 - P_{ii}), \tag{3.10}$$

for all $i, i_0 = 1, 2, \cdots, N$ and $j, j_0 = 0, 1, \cdots$.

Let

$$\pi^* = (\pi_1(1-P_{11}), \cdots, \pi_N(1-P_{NN}), \cdots, \pi_1 P_{11}^j(1-P_{11}), \cdots, \pi_N(1-P_{NN})P_{NN}^j, \cdots),$$

where π^*_{Nj+i}, the $(Nj+i)$-th element of π^*, is equal to $\pi_i P_{ii}^j(1-P_{ii})$ for all $i = 1, 2, \cdots, N$ and $j = 0, 1, \cdots$. We can verify that the limiting distribution derived in the above theorem is the stationary distribution of $\{\alpha_n\}$, which is defined in (2.10). More explicitly, $\pi^* Q = \pi^*$, where Q is the transition matrix for $\{\alpha_n\}$ defined in (2.11).

Not only we can show that $\{\alpha_n^0\}$ is a two-dimensional Markov chain with limiting distributions, we can also show the same property for the truncated hidden Markov chain $\alpha_n^{(D,0)}$, as defined in (2.12):

Theorem 3.7. *$\{\alpha_n^{(D,0)}\}$ is a two-dimensional Markov chain with the following limiting distribution: for each $i, i_0 = 1, \cdots, N$ and $j, j_0 = 0, \cdots, D$,*

$$\lim_{n\to\infty} P(S_n = i, \beta_n^{(D)} = j | S_0 = i_0, \beta_0^{(D)} = j_0) = \pi_i P_{ii}^j (1-P_{ii})^{I_{\{j<D\}}}, \qquad (3.11)$$

where $I_{\{j<D\}}$ is an indicator function, and $\beta_n^{(D)}$ is defined in (2.12).

We can also show that $\alpha_n^{(D,0)}$ converges in distribution to $\{\alpha_n^0\}$ as $D \to \infty$. In fact, for each $i \in \mathbb{S}$ and $j = 0, 1, \cdots$, when $D > j$, the definition of $\beta_n^{(D)}$ implies $P(S_n = i, \beta_n^{(D)} = j) = P(S_n = i, \beta_n = j)$. Therefore, it follows immediately that

$$\lim_{D\to\infty} P(S_n = i, \beta_n^{(D)} = j) = P(S_n = i, \beta_n = j).$$

This connection between $\{\alpha_n^{(D,0)}\}$ and $\{\alpha_n^0\}$ justifies the approximation of $\{\alpha_n^0\}$ by $\{\alpha_n^{(D,0)}\}$ in Chapter 2.

3.4 TDRS General Vasicek Model

In this section, we discuss stationarity of the TDRS General Vasicek model (2.4) and the properties of its first two moments. We would like to point out that proofs

for Lemma 3.2 and 3.3, and Theorem 3.9 are based on the ideas developed in Mao and Yuan (2006).

3.4.1 Stationarity of the Process

The stationarity we study in this section is "asymptotic stability in distribution" defined below (Mao and Yuan, 2006). In the rest of this chapter, we will be using stability and stationarity to describe the same property of the process.

Definition 3.6. *The process $(V(t), Y(t))$ is said to be asymptoticaly stable in distribution if there exists a probability measure $\pi(\cdot \times \cdot)$ on $\mathbb{R}^2$ such that the transition probability $p(t, v_0, y_0, dv \times dy)$ of $(V(t), Y(t))$ converges weakly to $\pi(dv \times dy)$ as $t \to \infty$ for every $V(0) = v_0, Y(0) = y_0$.*

Consider the TDRS General Vasicek model presented in (2.4):

$$dX(t) = a(\theta(S(t), \beta(t)) - X(t))dt + \sigma dW(t).$$

Then $\{X(t)\}$ is a continuous semi-martingale, provided that the conditions for the existence of a solution are satisfied (Theorem 3.4). We can present the solution to this SDE as a sum of two processes:

$$X(t) = I(t) + Z(t), \tag{3.12}$$

where

$$I(t) = I(0)e^{-at} + a\int_0^t e^{-a(t-u)}\theta(S(u), \beta(u))du, \tag{3.13}$$

and

$$Z(t) = e^{-at}Z(0) + \sigma\int_0^t e^{-a(t-s)}dW(s). \tag{3.14}$$

This decomposition can be verified by applying Itô's lemma to $H(t) = e^{at}X(t)$:

$$\begin{aligned} dH(t) &= e^{at}dX(t) + ae^{at}X(t)dt \\ &= e^{at}a\theta(S(t),\beta(t))dt - aX(t)e^{at}dt + \sigma e^{at}dW(t) + aX(t)e^{at}dt \\ &= e^{at}a\theta(S(t),\beta(t))dt + \sigma e^{at}dW(t). \end{aligned}$$

Integrating both sides, we get

$$H(t) = H(0) + \sigma\int_0^t e^{au}dW(u) + a\int_0^t e^{au}\theta(S(u),\beta(u))du.$$

Therefore,

$$\begin{aligned} X(t) &= e^{-at}X(0) + \sigma\int_0^t e^{-a(t-u)}dW(u) + a\int_0^t e^{-a(t-u)}\theta(S(u),\beta(u))du \qquad (3.15) \\ &= e^{-at}Z(0) + \sigma\int_0^t e^{-a(t-u)}dW(u) + e^{-at}I(0) + a\int_0^t e^{-a(t-u)}\theta(S(u),\beta(u))du \\ &= Z(t) + I(t), \end{aligned}$$

where $X(0) = Z(0) + I(0)$. Moreover, the following results can be verified

$$\begin{aligned} dZ(t) &= -aZ(t)dt + \sigma dW(t) \\ dI(t) &= a\left(\theta(S(t),\beta(t)) - I(t)\right)dt. \end{aligned}$$

Since by our assumption $\{Z(t)\}$ and $\{I(t)\}$ are independent, the stationarity of $\{X(t)\}$ can be established if both $\{Z(t)\}$ and $\{I(t)\}$ are asymptotically stable in distribution. Note that $\{Z(t)\}$ is an Ornstein-Uhlenbeck process and hence asymptotically stable in distribution. From now on, we focus on properties of the limiting distribution of $I(t)$. Stationarity of this process follows from stationarity of the two-dimensional process $(I(t), \theta(S(t),\beta(t)))$. To prove that the latter is stationary, we show asymptotic stability for a two-dimensional process $(V(t), Y(t))$, which is more general than $(I(t), \theta(S(t),\beta(t)))$. The two-dimensional process $(V(t), Y(t))$ is defined as follows:

$$\begin{aligned} V(t) &= e^{-at}V(0) + ae^{-at}\int_0^t e^{au}Y(u)du \\ Y(u) &= f(S(u),\beta(u)), \qquad (3.16) \end{aligned}$$

where f is an injective function of $\alpha_u = (S(u), \beta(u))'$. We assume that for each $i \in \mathbb{S}$, $f(i,t)$ is a continuous function of t. We also assume that Θ, the range for $\{Y(t)\}$, is uniformly bounded by M for some $M > 0$ and all $t > 0$. These assumptions are all satisfied by the TDRS Vasicek model. Let $V^{v_0,y_0}(t)$ denote a solution $(V(t), Y(t))$ to (3.16) corresponding to the initial value $V(0) = v_0, Y(0) = y_0$. From Corollary 3.1, $\{Y(t)\}$ is a homogeneous Markov process. For the two-dimensional process $\{V(t), Y(t)\}$, we have the following result:

Theorem 3.8. *The process $\{V(t), Y(t)\}$ is a two-dimensional homogeneous strong Markov process.*

Now we state the main result of this section:

Theorem 3.9. *The process $\{V(t), Y(t)\}$ defined in (3.16) is asymptotically stable in distribution.*

The proof of the above theorem is based on several lemmas. The connections between the lemmas and our main results are as follows: Lemmas 3.2 and 3.3 are used in the proof of Theorem 3.9; Lemma 3.2 is used for proof of Lemma 3.3; Lemma 3.1 is the one used to prove Lemma 3.2. Before stating Lemma 3.1, we introduce several notations. For $1 \leq i_1, i_2 \leq N$ and $t_1 > 0, t_2 > 0$, let the process $\{S_{i_1}(t)\}_{t\geq 0}$ be $\{S(t)\}_{t\geq 0}$ starting with $S_0 = i_1$ and the process $\{\beta_{t_1}(t)\}_{t\geq 0}$ be $\{\beta(t)\}_{t\geq 0}$ starting with $\beta_0 = t_1$. Likewise we have $\{S_{i_2}(t)\}_{t\geq 0}$ and $\{\beta_{t_2}(t)\}_{t\geq 0}$. Using the idea of coupling (Thorisson, 2000), we can impose that $\{S_{i_2}(t)\}$ be the same as $\{S_{i_1}(t)\}$ after $\{S_{i_1}(t)\}$ and $\{S_{i_2}(t)\}$ meet for the first time. The first meeting time is finite with probability one (Anderson, 1991).

Lemma 3.1. *Define*

$$T^{(t_1,t_2)}_{(i_1,i_2)} = \inf\{t \geq 0 : S_{i_1}(t) = S_{i_2}(t), \beta_{t_1}(t) = \beta_{t_2}(t)\}.$$

Then for each ϵ there exists $T > 0$ such that

$$P(T^{(t_1,t_2)}_{(i_1,i_2)} > T) < \epsilon.$$

By the one-to-one correspondence between $\{Y(t)\}$ and $(S(t), \beta(t))$, the above lemma implies that $\{Y(t)\}$ starting with different values meet together after a finite time almost surely. It follows from the definition of stopping time that $T_{(i_1,i_2)}^{(t_1,t_2)}$ is a stopping time. In Lemma 3.2, we make use of the strong Markov property of $(V(t), Y(t))$ to prove uniform convergence results of transition probabilities of the process $(V(t), Y(t))$ with respect to different initial values.

Now, let $\mathcal{P}(\mathbb{R} \times \Theta)$ denote the set of all probability measures on $\mathbb{R} \times \Theta$, where Θ is the range of $Y(t)$. For $P_1, P_2 \in \mathcal{P}(\mathbb{R} \times \Theta)$, define bounded Lipschitz metric $d_{\mathbb{L}}$ as follows (Definition 3.2):

$$d_{\mathbb{L}}(P_1, P_2) = \sup_{f \in \mathbb{L}} \left| \int_{\Theta} \int_{\mathbb{R}} f(v, y) P_1(dv, dy) - \int_{\Theta} \int_{\mathbb{R}} f(v, y) P_2(dv, dy) \right|,$$

where

$$\mathbb{L} = \{f : \mathbb{R} \times \Theta \to \mathbb{R} : |f(v_2, y_2) - f(v_1, y_1)| \leq |v_2 - v_1| + |y_2 - y_1| \text{ and } |f(\cdot, \cdot)| \leq 1\}.$$

Lemma 3.2. *Let $p(t, v, y, dv \times dy)$ denote the transition probability of the process $(V(t), Y(t))$. For any compact subset K of $\mathbb{R}$,*

$$\lim_{t \to \infty} d_{\mathbb{L}}(p(t, v_2, y_2, \cdot \times \cdot), p(t, v_1, y_1, \cdot \times \cdot)) = 0,$$

uniformly for $v_1, v_2 \in K$, and $y_1, y_2 \in \Theta$.

The following lemma states that $\{p(t, v_0, y_0, \cdot \times \cdot) : t \geq 0\}$ is a Cauchy sequence in $\mathcal{P}(\mathbb{R} \times \Theta)$.

Lemma 3.3. *For any $(v_0, y_0) \in \mathbb{R} \times \Theta$, $\{p(t, v_0, y_0, \cdot \times \cdot) : t \geq 0\}$ is Cauchy in the space $\mathcal{P}(\mathbb{R} \times \Theta)$ with metric $d_{\mathbb{L}}$.*

It follows from the above lemma and the definition of completeness of the bounded Lipschitz metric that there exists a probability measure $\pi(\cdot \times \cdot) \in \mathcal{P}(\mathbb{R} \times \Theta)$ such that, for any $(v_0, y_0) \in \mathbb{R} \times \Theta$, the transition probabilities $\{p(t, v, y, \cdot \times \cdot) : t \geq 0\}$ converge to $\pi(\cdot \times \cdot)$, which is the claim of Theorem 3.9.

The following result states that the first two moments of $I(t)$ converge to constants as $t \to \infty$.

Corollary 3.2. *Let $I(t)$ be as defined in (3.13). Then both $\lim_{t\to\infty} E(I(t))$ and $\lim_{t\to\infty} Var(I(t))$ exist.*

From (3.12), we know that $X(t) = Z(t) + I(t)$. Since $Z(t)$ is an Ornstein-Uhlenbeck process, the first two moments of $Z(t)$ converge to constants as $t \to \infty$. From Corollary 3.2 the first two moments of $I(t)$ also converge to constants as $t \to \infty$. Due to independence of $Z(t)$ and $I(t)$, it follows that the first two moments of $X(t)$ converge to constants as well.

3.5 Concluding Remarks

In this chapter, we have proved the existence and uniqueness of the solution to our proposed TDRS model. We have also proved the stationarity of the TDRS General Vasicek model, which is a special case of TDRS model and depends on the explicit form of a solution to the TDRS General Vasicek model (2.4). One natural extension of our results would be to prove the following conjecture: the TDRS model (2.1) is stationary for more general forms of the drift function f.

3.6 Appendix: Technical Proofs

Proof of Theorem 3.4

By the definition of a continuous-time Markov chain, almost every sample path of $S(\cdot)$ is a right-continuous step function with a finite number of jumps on any finite interval $[t_0, T]$. Let $\tau_0 = t_0 - \beta_0$, which is $\mathcal{F}_{t_0}$ measurable. We define a

sequence of stopping times:

$$\tau_{i+1} = \inf\{\tau_i < t \leq T : S(t) \neq S(\tau_i)\}, \quad i \in \mathbb{Z}^+.$$

It is clear that $\tau_i = T$ for sufficiently large i. Define $\hat{\mathcal{F}}(t) = \mathcal{F}_{t+t_0}$ and $\hat{W}(t) = W(t+t_0)$. Consider the following equation:

$$d\hat{X}(t) = f(\hat{X}(t), S(t_0), t)dt + g(\hat{X}(t), S(t_0), t)d\hat{W}(t), \quad t \in [0, T-t_0] \tag{3.17}$$

with initial values $S(t_0)$ and $\hat{X}_0$. By Theorem 3.2 and its remarks, there exists a unique solution $\{\hat{X}(t)\}$ to (3.17) which belongs to $\mathcal{M}^2([0, \tau_1 - t_0]; \mathbb{R})$. In particular, $\hat{X}(\tau_1 - t_0) \in \mathcal{L}^2_{\mathcal{F}_{\tau_1}}(\Omega; \mathbb{R})$. If we take $X(t) = \hat{X}(t - t_0)$ for $t \in [t_0, \tau_1]$, then $\{X(t)\}$ is a solution to (2.1) on $[t_0, \tau_1]$. Next, we define $\hat{\mathcal{F}}(t) = \mathcal{F}_{t+\tau_1}$, $\hat{W}(t) = W_{t+\tau_1}$, and $\hat{X}_0 = X(\tau_1)$. Consider the following equation:

$$d\hat{X}(t) = f(\hat{X}(t), S(\tau_1), t)dt + g(\hat{X}(t), S(\tau_1), t)d\hat{W}(t), \quad t \in [0, T-\tau_1] \tag{3.18}$$

with the initial values $S(\tau_1)$ and $\hat{X}_0$. Again, Theorem 3.2 and its remarks imply that there exists a unique solution $\{\hat{X}(t)\}$ to (3.18), which belongs to $\mathcal{M}^2([0, \tau_2 - \tau_1]; \mathbb{R})$. In particular, $\hat{X}_{\tau_2 - \tau_1} \in \mathcal{L}^2_{\mathcal{F}_{\tau_2}}(\Omega; \mathbb{R})$. For $X(t) = \hat{X}_{t-\tau_1}$ for $t \in [\tau_1, \tau_2]$, the process $\{X(t)\}$ is a solution to (2.1) on $[\tau_1, \tau_2]$. Repeating the above procedure, we can find a unique solution $X(t)$ to (2.1) on $[t_0, T]$.

Using the same approach as in the proof of Lemma 3.1 in Mao and Yuan (2006), we can also show (3.9). □

Proof of Theorem 3.5

To show the Markov property of $\{\alpha(t)\}$, let A be a Borel set in $\mathbb{R}$ and $h < t$.

Then,

$$\begin{aligned}
& P(\beta(t) \in A, S(t) = i | S(u), \beta(u), u \in [0,h]) \\
= \; & P(\beta(t) \in A | S(t) = i, S(u), \beta(u), u \in [0,h]) P(S(t) = i | S(u), \beta(u), u \in [0,h]) \\
= \; & P(\beta(t) \in A | S(t) = i, S(h), \beta(h)) P(S(t) = i | S(h), \beta(h)), \text{by Markov property} \\
= \; & P(\beta(t) \in A, S(t) = i | S(h), \beta(h)).
\end{aligned}$$

To show the limiting distribution of $\{\alpha(t)\}$, for each $x \in \mathbb{R}^+$ and $i \in \mathbb{S}$, we assume $t > h + x$. Then we have

$$\begin{aligned}
& P(\beta(t) \geq x, S(t) = i | S(u), \beta(u), u \in [0,h]) \\
= \; & P(S_v = i, \forall v \in [t-x, t] | S(u), \beta(u), u \in [0,h]) \\
= \; & P(S_v = i, \forall v \in [t-x, t] | S_{t-x} = i, S(u), \beta(u), u \in [0,h]) \\
& \times P(S_{t-x} = i | S(u), \beta(u), u \in [0,h]) \\
= \; & P(S_v = i, \forall v \in [t-x, t] | S_{t-x} = i) P(S_{t-x} = i | S(u), \beta(u), u \in [0,h]) \\
= \; & \exp(\gamma_{ii} x) P(S_{t-x} = i | S_h) \\
\longrightarrow \; & \pi_i \exp(\gamma_{ii} x) \text{ as } t \to \infty, \qquad (3.19)
\end{aligned}$$

where the third equality follows from the Markov property, the fourth equality follows from the exponential distribution of the waiting time and Markov property, and the last equality follows from the limiting distribution of $\{S(t)\}$. From (3.19), it follows that

$$P(\beta(t) \leq x, S(t) = i | S(u), \beta(u), u \in [0,h]) \stackrel{t \to \infty}{\longrightarrow} \pi_i (1 - \exp(\gamma_{ii} x)).$$

Finally, the homogeneity of $\{\alpha(t)\}$ follows from that of $\{S(t)\}$. □

Proof of Corollary 3.1

The result follows from the Markov property and homogeneity of $\alpha(t) = (S(t), \beta(t))$ that are established in Theorem 3.5. □

Proof of Theorem 3.6

To prove the Markov property of α_n^0, it is sufficient to prove that for each $i, i_k \in \mathbb{S}$ and $j, j_k = 0, 1, \cdots,$

$$\begin{aligned} & P(S_n = i, \beta_n = j | S_{n-1} = i_{n-1}, \beta_{n-1} = j_{n-1}, \cdots, S_0 = i_0, \beta_0 = j_0) \\ = \; & P(S_n = i, \beta_n = j | S_{n-1} = i_{n-1}, \beta_{n-1} = j_{n-1}). \end{aligned}$$

Indeed,

$$\begin{aligned} & P(S_n = i, \beta_n = j | S_{n-1} = i_{n-1}, \beta_{n-1} = j_{n-1}, \cdots, S_0 = i_0, \beta_0 = j_0) \\ = \; & P(\beta_n = j | S_n = i, S_{n-1} = i_{n-1}, \beta_{n-1} = j_{n-1}, \cdots, S_0 = i_0, \beta_0 = j_0) \\ & \times P(S_n = i | S_{n-1} = i_{n-1}, \beta_{n-1} = j_{n-1}, \cdots, S_0 = i_0, \beta_0 = j_0) \\ = \; & P(\beta_n = j | S_n = i, S_{n-1} = i_{n-1}, \beta_{n-1} = j_{n-1}) P(S_n = i | S_{n-1} = i_{n-1}), \end{aligned}$$

where the last equality holds because β_n can be uniquely determined by the values of S_n, S_{n-1} and β_{n-1}, and $\{S_n\}$ has Markov property. The Markov property of $\{S_n\}$ also implies that

$$P(S_n = i | S_{n-1} = i_{n-1}, \beta_{n-1} = j_{n-1}) = P(S_n = i | S_{n-1} = i_{n-1}).$$

Therefore,

$$\begin{aligned} & P(\beta_n = j | S_n = i, S_{n-1} = i_{n-1}, \beta_{n-1} = j_{n-1}) P(S_n = i | S_{n-1} = i_{n-1}) \\ = \; & P(\beta_n = j | S_n = i, S_{n-1} = i_{n-1}, \beta_{n-1} = j_{n-1}) P(S_n = i | S_{n-1} = i_{n-1}, \beta_{n-1} = j_{n-1}) \\ = \; & P(S_n = i, \beta_n = j | S_{n-1} = i_{n-1}, \beta_{n-1} = j_{n-1}). \end{aligned}$$

To derive the limiting distribution of (S_n, β_n),

$$\begin{aligned}
& P(S_n = i, \beta_n \geq j | S_0 = i_0, \beta_0 = j_0) \\
= \quad & P(S_n = S_{n-1} = \cdots = S_{n-j} = i | S_0 = S_{-1} = \cdots = S_{-j_0} = i_0, S_{-j_0-1} \neq i_0) \\
= \quad & P(S_n = S_{n-1} = \cdots = S_{n-j} = i | S_{n-j} = i, S_0 = S_{-1} = \cdots = S_{-j_0} = i_0, S_{-j_0-1} \neq i_0) \\
& \times P(S_{n-j} = i | S_0 = S_{-1} = \cdots = S_{-j_0} = i_0, S_{-j_0-1} \neq i_0) \\
= \quad & P_{ii}^j P(S_{n-j} = i | S_0 = i_0) \text{ by Markov property} \\
\longrightarrow \quad & \pi_i P_{ii}^j, \text{ as } n \to \infty, \text{ by the limiting distribution of } \{S_n\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& P(S_n = i, \beta_n = j | S_0 = i_0, \beta_0 = j_0) \\
= \quad & P(S_n = i, \beta_n \geq j | S_0 = i_0, \beta_0 = j_0) - P(S_n = i, \beta_n \geq j+1 | S_0 = i_0, \beta_0 = j_0) \\
\longrightarrow \quad & \pi_i P_{ii}^j - \pi_i P_{ii}^{j+1}, \text{ as } n \to \infty \\
= \quad & \pi_i P_{ii}^j (1 - P_{ii}).
\end{aligned}$$

□

Proof of Theorem 3.7

The proof is similar to the one presented for Theorem 3.6. Here we show only the limiting distribution of $\alpha_n^{(D)}$. For $j = 0, 1, \cdots D-1$, $\beta_n^{(D)} = \beta_n$ and hence

$$\begin{aligned}
& P(S_n = i, \beta_n^{(D)} = j | S_0 = i_0, \beta_0^{(D)} = j_0) \\
= \quad & P(S_n = i, \beta_n = j | S_0 = i_0, \beta_0^{(D)} = j_0) \\
\stackrel{n\to\infty}{\longrightarrow} \quad & \pi_i P_{ii}^j (1 - P_{ii}), \qquad (3.20)
\end{aligned}$$

where the convergence follows from (3.10). On the other hand, by the definition of

$\beta_n^{(D)}$,

$$\begin{aligned}
& P(S_n = i, \beta_n^{(D)} = D | S_0 = i_0, \beta_0^{(D)} = j_0) \\
= \quad & P(S_n = i, \beta_n \geq D | S_0 = i_0, \beta_0^{(D)} = j_0) \\
= \quad & P(S_n = i | S_0 = i_0, \beta_0^{(D)} = j_0) - \sum_{j=0}^{D-1} P(S_n = i, \beta_n = j | S_0 = i_0, \beta_0^{(D)} = j_0) \\
\stackrel{n\to\infty}{\longrightarrow} \quad & \pi_i - \pi_i \sum_{j=0}^{D-1} P_{ii}^j (1 - P_{ii}) = \pi_i P_{ii}^D,
\end{aligned}$$

where the convergence follows from (3.20). □

Proof of Theorem 3.8

$$\begin{aligned}
V(t+s) &= e^{-a(t+s)} V(0) + a e^{-a(t+s)} \int_0^{t+s} e^{au} Y(u) du \\
&= e^{-at} [a e^{-as} V(0) + a e^{-as} \int_0^s e^{au} Y(u) du + a e^{-as} \int_s^{s+t} e^{au} Y(u) du] \\
&= e^{-at} [V(s) + a e^{-as} \int_s^{s+t} e^{au} Y(u) du] \\
&= e^{-at} V(s) + a e^{-a(s+t)} \int_s^{s+t} e^{au} Y(u) du.
\end{aligned}$$

It follows from the above equation and Markov property of $Y(t)$ that

$$\begin{aligned}
& P(V(t+s) \in A, Y(t+s) \in B | \mathcal{F}_s) \\
= \quad & P(V(t+s) \in A, Y(t+s) \in B | V(s), Y(s)).
\end{aligned}$$

Therefore, $(V(t), Y(t))$ is two-dimensional Markov process. The conditions that guarantee a Markov process the strong Markov property are the right continuity of the sample paths plus the so-called Feller property. By our assumption that $f(i,t)$ is a continuous function of t for each $i \in \mathbb{S}$, $Y(t) = f(S(t), \beta(t))$ has right-continuous sample paths. The form (3.16) and the boundedness of $Y(t)$ imply that

the sample path of $V(t)$ is continuous. Therefore, we only need to verify the Feller property of $(V(t), Y(t))$: for any bounded continuous function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ and any $v \in \mathbb{R}, y \in \Theta, \lambda > 0$, the mapping

$$(v, y, s) \to E\varphi(V^{v,y}(s+\lambda), Y^y(s+\lambda))$$

is continuous. In fact, the Feller property follows from the facts that $V(t)$ is a continuous function of v_0 and t, $f(i,t)$ is a continuous function of t for each $i \in \mathbb{S}$, and Θ is bounded, in conjunction with an application of the bounded convergence theorem. Finally, the homogeneity of $\{V(t), Y(t)\}$ follows from that of $\{\alpha(t)\}$ and one-to-one correspondence between $\{\alpha(t)\}$ and $Y(t)$. □

Proof of Lemma 3.1

Note that $\{\beta(t)\}_{t\geq 0}$ keeps track of partial information of $\{S(t)\}_{t\geq 0}$ by recording how long $\{S(t)\}_{t\geq 0}$ has spent in the current regime. Conditional on knowing the path of $\{S(t)\}_{t\geq 0}$, there is no further randomness in $\{\beta(t)\}_{t\geq 0}$. Let

$$\begin{aligned} T_{i_1}^{i_2} &= \inf\{t \geq 0 : S_{i_1}(t) = S_{i_2}(t)\} \\ T^W &= T_{i_1}^{i_2} + \inf\{t > 0 : S_{i_1}(t + T_{i_1}^{i_2}) \neq S_{i_1}(T_{i_1}^{i_2})\}. \end{aligned}$$

By the above definitions and our assumption that $\{S_{i_2}(t)\}$ is the same as $\{S_{i_1}(t)\}$ after $\{S_{i_1}(t)\}$ and $\{S_{i_2}(t)\}$ meet for the first time, we have the following facts:

$$\begin{aligned} S_{i_1}(t) &= S_{i_2}(t),\ \forall t > T_{i_1}^{i_2} \\ S_{i_1}(T^W) &= S_{i_2}(T^W) \\ \beta_{t_1}(T^W) &= \beta_{t_2}(T^W) = 0,\ \forall t_1, t_2 > 0. \end{aligned}$$

Therefore,

$$T_{(i_1,i_2)}^{(t_1,t_2)} \leq T^W, a.s.$$

Now,

$$\begin{aligned} & P(T_{(i_1,i_2)}^{(t_1,t_2)} > T) \\ \leq\ & P(T^W > T) \\ \leq\ & P(T_{i_1}^{i_2} > \frac{T}{2}) + P(\inf\{t > 0 : S_{i_1}(t + T_{i_1}^{i_2}) \neq S_{i_1}(T_{i_1}^{i_2})\} > \frac{T}{2}) \\ \leq\ & P(T_{i_1}^{i_2} > \frac{T}{2}) + \exp(\max_{1\leq i\leq N}\{-\gamma_{ii}\}T/2), \end{aligned}$$

where $\Gamma = (\gamma_{ij})$ is the infinitesimal generator of $S(t)$. By the ergodicity of $S(t)$, there exists T^* such that

$$P(T_{i_1}^{i_2} > \frac{T^*}{2}) < \frac{\epsilon}{2}, \quad \forall 1 \leq i_1, i_2 \leq N.$$

[Mao and Yuan, 2006; Anderson, 1991]. On the other hand, there exists $T^{**} > 0$, such that $\exp(\max_{1\leq i\leq N}\{-\gamma_{ii}\}T^{**}/2) < \frac{\epsilon}{2}$. Therefore, for $T > \max\{T^*, T^{**}\}$,

$$P(T_{(i_1,i_2)}^{(t_1,t_2)} > T) < \epsilon, \quad \forall 1 \leq i_1, i_2 \leq N \text{ and } t_1 > 0, t_2 > 0.$$

□

Proof of Lemma 3.2

Define the stopping time

$$T_{y_1}^{y_2} = \inf\{t \geq 0 : Y_{y_1}(t) = Y_{y_2}(t)\},$$

where $Y_y(t)$ is the process $Y(t)$ started at the initial value y. Since $Y(t) = f(S(t), \beta(t))$ is an injective function as defined in (3.16), for each $y \in \Theta$, there is a one-to-one correspondence between the value of $Y(t)$ and the pair $(S(t), \beta(t))$. It follows from Lemma 3.1 that $T_{y_1}^{y_2} < \infty, a.s.$ Moreover, for every $\epsilon > 0$, there exists a positive number T such that

$$P(T_{y_1}^{y_2} < T) > 1 - \frac{\epsilon}{4}, \quad \forall\ y_1, y_2 \in \Theta. \tag{3.21}$$

Since K is a compact subset of $\mathbb{R}$, we can assume $K \subseteq B(0,R)$, a ball centered at 0 with radius R. Let

$$T_1 = [\log(\frac{4R+8M}{\epsilon})] + 1,$$

where $[\cdot]$ means a rounded integer. For $t > T + T_1$ and $f \in \mathbb{L}$,

$$\begin{aligned}
& |Ef(V^{v_2,y_2}(t), Y_{y_2}(t)) - Ef(V^{v_1,y_1}(t), Y_{y_1}(t))| \\
= \; & |E(I_{T_{y_1}^{y_2} \geq T}(f(V^{v_2,y_2}(t), Y_{y_2}(t)) - f(V^{v_1,y_1}(t), Y_{y_1}(t))) \\
& + E(I_{T_{y_1}^{y_2} < T}(f(V^{v_2,y_2}(t), Y_{y_2}(t)) - f(V^{v_1,y_1}(t), Y_{y_1}(t))))| \\
\leq \; & 2P\{T_{y_1}^{y_2} \geq T\} + E(I_{T_{y_1}^{y_2} < T}|f(V^{v_2,y_2}(t), Y_{y_2}(t)) - f(V^{v_1,y_1}(t), Y_{y_1}(t))|) \\
= \; & 2P\{T_{y_1}^{y_2} \geq T\} \\
& + E\{I_{T_{y_1}^{y_2} < T} E(|f(V^{v_2,y_2}(t), Y_{y_2}(t)) - f(V^{v_1,y_1}(t), Y_{y_1}(t))| \mathcal{F}_{T_{y_1}^{y_2}}) \\
\leq \; & \frac{\epsilon}{2} + E(|f(V^{v_2,y_2}(t), Y_{y_2}(t)) - f(V^{v_1,y_1}(t), Y_{y_1}(t))| \Big| \mathcal{F}_{T_{y_1}^{y_2}})\}, \qquad (3.22)
\end{aligned}$$

where $I_{(\cdot)}$ is the indicator function and the last equality follows from (3.21). Then, from (3.22), the strong Markov property of the process $\{V(t), Y(t)\}$ and the definition of $\mathbb{L}$,

$$\begin{aligned}
& |Ef(V^{v_2,y_2}(t), Y_{y_2}(t)) - Ef(V^{v_1,y_1}(t), Y_{y_1}(t))| \\
\leq \; & \frac{\epsilon}{2} + E\{I_{T_{y_1}^{y_2} < T} E(|f(V^{u_2,z}(t - T_{y_1}^{y_2}), Y_z(t - T_{y_1}^{y_2})) - f(V^{u_1,z}(t - T_{y_1}^{y_2}), Y_z(t - T_{y_1}^{y_2}))|)\} \\
\leq \; & \frac{\epsilon}{2} + E\{I_{T_{y_1}^{y_2} < T} E(2 \wedge |V^{u_2,z}(t - T_{y_1}^{y_2}) - V^{u_1,z}(t - T_{y_1}^{y_2})|)\}, \qquad (3.23)
\end{aligned}$$

where

$$u_2 = V^{v_2,y_2}(T_{y_1}^{y_2}), \quad u_1 = V^{v_1,y_1}(T_{y_1}^{y_2}), \quad z = Y_{y_2}(T_{y_1}^{y_2}) = Y_{y_1}(T_{y_1}^{y_2}).$$

By definition of $V(t)$ in (3.16),

$$|V^{u_2,z}(t - T_{y_1}^{y_2}) - V^{u_1,z}(t - T_{y_1}^{y_2})| = |e^{-a(t - T_{y_1}^{y_2})}(u_2 - u_1)|. \qquad (3.24)$$

Note that

$$\begin{aligned} |u_1| &= |V^{v_1,y_1}(T_{y_1}^{y_2})| \\ &= |e^{-aT_{y_1}^{y_2}} v_1 + ae^{-aT_{y_1}^{y_2}} \int_0^{T_{y_1}^{y_2}} e^{au} Y(u)du| \\ &\leq e^{-aT_{y_1}^{y_2}} |v_1| + ae^{-aT_{y_1}^{y_2}} M \frac{e^{aT_{y_1}^{y_2}} - 1}{a} \\ &\leq R + 2M. \end{aligned}$$

Similarly,

$$|u_2| \leq R + 2M.$$

From the above results and (3.24), we have

$$\begin{aligned} |V^{u_2,z}(t - T_{y_1}^{y_2}) - V^{u_1,z}(t - T_{y_1}^{y_2})| &\leq (2R + 4M)e^{-a(t - T_{y_1}^{y_2})} \\ &\leq (2R + 4M)e^{-aT_1} < \frac{\epsilon}{2}, \end{aligned} \tag{3.25}$$

where the last inequality follows from the definition of T_1. Therefore, it follows from (3.23) and (3.25) that

$$|Ef(V^{v_2,y_2}(t), Y_{y_2}(t)) - Ef(V^{v_1,y_1}(t), Y_{y_1}(t))| < \epsilon.$$

Due to the arbitrariness of ϵ and $f \in \mathbb{L}$, we have proved our claim. □

Proof of Lemma 3.3

By the definition of Cauchy sequence, for any $(v_0, y_0) \in \mathbb{R} \times \Theta$, we need to show only that for every $\epsilon > 0$, there exists $T > 0$ such that

$$d_{\mathbb{L}}(p(t + s, v_0, y_0, \cdot \times \cdot), p(t, v_0, y_0, \cdot \times \cdot)) \leq \epsilon, \text{ for every } t \geq T, s > 0,$$

which is equivalent to

$$\sup_{f \in \mathbb{L}} |Ef(V^{v_0,y_0}(t + s), Y_{y_0}(t + s)) - Ef(V^{v_0,y_0}(t), Y_{y_0}(t))| \leq \epsilon, \textit{for any } t \geq T, s > 0.$$

Now, for $f \in \mathbb{L}$, $t > 0, s > 0$,

$$\begin{aligned}
& |Ef(V^{v_0,y_0}(t+s), Y_{y_0}(t+s)) - Ef(V^{v_0,y_0}(t), Y_{y_0}(t))| \\
= \; & |E[E(f(V^{v_0,y_0}(t+s), Y_{y_0}(t+s))|\mathcal{F}_s)] - Ef(V^{v_0,y_0}(t), Y_{y_0}(t))| \\
= \; & |\int_\Theta \int_{\mathbb{R}} Ef(V^{z,l}(t), Y_l(t)) p(s, v_0, y_0, dz \times dl) - Ef(V^{v_0,y_0}(t), Y_{y_0}(t))| \\
\leq \; & \int_\Theta \int_{\mathbb{R}} |Ef(V^{z,l}(t), Y_l(t)) - Ef(V^{v_0,y_0}(t), Y_{y_0}(t))| p(s, v_0, y_0, dz \times dl),
\end{aligned}$$

where $z = V^{v_0,y_0}(s), l = Y_{y_0}(s)$ and the second equality follows from homogeneity and Markov property of $(V(t), Y(t))$. Note that by definition of $V(t)$ in (3.16) and boundedness of $Y(t)$,

$$\begin{aligned}
|V^{v_0,y_0}(t)| &\leq e^{-at}|v_0| + aMe^{-at}\int_0^t e^{au}du & (3.26) \\
&\leq |v_0| + M, & (3.27)
\end{aligned}$$

where M is an upper bound for $Y(t)$. Therefore, $|z| \leq |v_0| + M$. By Lemma 3.2,

$$\lim_{t\to\infty} \sup_{f\in\mathbb{L}} |Ef(V^{z,l}(t+s), Y_l(t)) - Ef(V^{v_0,y_0}(t), Y_{y_0}(t))| = 0,$$

uniformly for $z = V^{v_0,y_0}(s)$. Then it follows that

$$\lim_{t\to\infty} d_{\mathbb{L}}(p(t+s, v_0, y_0, \cdot\times\cdot), p(t, v_0, y_0, \cdot\times\cdot)) = 0.$$

□

Proof of Theorem 3.9

We need to show that there exists a probability measure $\pi(\cdot\times\cdot) \in \mathcal{P}(\mathbb{R}\times\Theta)$ such that, for any $(v, y) \in \mathbb{R}\times\Theta$, the transition probabilities $\{p(t, v, y, \cdot\times\cdot) : t \geq 0\}$ converge weakly to $\pi(\cdot\times\cdot)$. By Theorem 3.1, we need to show only the convergence results under the bounded Lipschitz metric. For initial values v_0, y_0, we know by Lemma 3.3 that $\{p(t, v_0, y_0, \cdot\times\cdot) : t \geq 0\}$ is Cauchy in the space $\mathcal{P}(\mathbb{R}\times\Theta)$. Therefore,

there exists a unique $\pi(\cdot\times\cdot)$ in $\mathcal{P}(\mathbb{R}\times\Theta)$ such that $\{p(t,v_0,y_0,\cdot\times\cdot):t\geq 0\}$ converge weakly to $\pi(\cdot\times\cdot)$, i.e.

$$\lim_{t\to\infty} d_{\mathbb{L}}(p(t,v_0,y_0,\cdot\times\cdot),\pi(\cdot\times\cdot))=0. \tag{3.28}$$

Now we prove that for any other initial values (v,y), $\{p(t,v,y,\cdot\times\cdot):t\geq 0\}$ converge weakly to the same limit $\pi(\cdot\times\cdot)$. It follows from Lemma 3.2 that

$$\lim_{t\to\infty} d_{\mathbb{L}}(p(t,v,y,\cdot\times\cdot),p(t,v_0,y_0,\cdot\times\cdot))=0.$$

The above result in conjunction with (3.28) implies that

$$\begin{aligned}
& \lim_{t\to\infty} d_{\mathbb{L}}(p(t,v,y,\cdot\times\cdot),\pi(\cdot\times\cdot)) \\
\leq\ & \lim_{t\to\infty}[d_{\mathbb{L}}(p(t,v,y,\cdot\times\cdot),p(t,v_0,y_0,\cdot\times\cdot))+d_{\mathbb{L}}(p(t,v_0,y_0,\cdot\times\cdot),\pi(\cdot\times\cdot))] \\
=\ & 0.
\end{aligned}$$

□

Proof of Corollary 3.2

It follows from Theorem 3.9 that the vector process $(V(t),Y(t))$ defined in (3.16) is asymptotically stable in distribution. With $Y(t)$ replaced by $\theta(S(t),\beta(t))$, $V(t)$ becomes $I(t)$ as defined in (3.13). Therefore, $(I(t),\theta(S(t),\beta(t))$ is asymptotically stable in distribution.

By Definition 3.6, the transition probability of $(I(t),\theta(S(t),\beta(t)))$ converges weakly to a certain probability measure $\pi(\cdot\times\cdot)$ on $\mathbb{R}^2$. Alternatively, for any bounded real-valued continuous function $f(\cdot,\cdot)$ on $\mathbb{R}^2$, $E[f(I(t),\theta(S(t),\beta(t)))]$ converges to $E[f(\xi,\eta)]$, where (ξ,η) are random variables with joint probability density function $\pi(\cdot\times\cdot)$. In the following, we prove that both $\lim_{t\to\infty} E(I(t))$ and $\lim_{t\to\infty} Var(I(t))$ exist by choosing a particular $f(\cdot,\cdot)$.

By definition in (2.5), $\theta(S(t),\beta(t))$ is uniformly bounded for $t>0$. It follows from the form (3.13) that $I(t)$ is also uniformly bounded for $t>0$. Without loss

of generality, suppose that $|I(t)| < M$ for all $t > 0$. Define $f(x, y) = x$ for $|x| < M$ and $f(x, y) = M$ for $|x| \geq M$. Then $f(x, y)$ is a bounded real-valued continuous function on $\mathbb{R}^2$, and for each $t > 0$, $f(I(t), \theta(S(t), \beta(t))) = I(t)$. Therefore, from our previous conclusion, $E[f(I(t), \theta(S(t), \beta(t)))] = E[I(t)]$ converges to $E[f(\xi, \eta)]$ as $t \to \infty$, where (ξ, η) are random variables with joint probability density function $\pi(\cdot \times \cdot)$. In other words, $E[I(t)]$ converges to a constant when $t \to \infty$. We can prove that $\lim_{t\to\infty} E[I^2(t)]$ exists in a similar manner. Then it follows that $\lim_{t\to\infty} Var(I(t))$ exists as well. $\square$

Chapter 4

Inference for Time-Dependent Drift

4.1 Introduction

In this chapter, we study methods of estimating a time-dependent drift component for the following class of SDEs:

$$dr(t) = \theta(t)dt + \sigma(t, r(t))dW(t), \quad 0 \leq t \leq T. \tag{4.1}$$

For the rest of this chapter, we refer to $\theta(t)$ as the drift function. In the case of discrete observations of the process, the change in the values of the process between observation times are impacted by both the drift and diffusion components. Due to this confounding effect between the drift and diffusion terms, identification of the former from discrete observations is generally not possible. The problem can be resolved in the context of stationary processes. For example, Jiang and Knight (1997) study a strictly stationary process of the following form:

$$dr(t) = \mu(r(t))dt + \sigma(r(t))dW(t), \quad 0 \leq t \leq T.$$

The authors propose nonparametric estimators of both the drift function $\mu(\cdot)$ and diffusion function $\sigma(\cdot)$. They estimate the diffusion function first and then estimate the drift function based on the diffusion function estimator and the marginal density of the process. The nonparametric estimator proposed therein is asymptotically consistent when the sampling interval shrinks to zero and the trajectory length of the process tends to infinity. In our case, the process (4.1) is assumed to be non-stationary, and hence the argument used by the authors does not apply. To simplify the analysis, we assume that a continuous realization of the process $\{r(t)\}$ is available. This assumption implies that, for each t, the diffusion function value $\{\sigma(t, r(t))\}$ is completely known, as otherwise, it can be estimated without error from the quadratic variation of the process.

We approach the problem of estimating $\theta(t)$ by applying maximum likelihood together with the sieve method. The likelihood function is maximized over a parameter space spanned by a sequence of polynomial functions. In other words, we estimate the projection of $\theta(t)$ onto a finite dimensional space, and the parameters are defined to be the coefficients in the projection. Our estimation method is different from some well-known nonparametric estimation approaches proposed in the literature, such as Kalman filter and moving average. Since our projection space is spanned by polynomial functions, the resulting estimator of $\theta(t)$ is a smooth function. However, the Kalman filter results in a non-smooth estimator because it uses conditional expectation of a diffusion process. Moving average estimator is typically used in the context of time series (e.g. Hyndman, 2008).

Since we always observe financial or economic data over a finite time interval, we are not able to estimate too many parameters at a given level of accuracy. In particular, when $\theta(t)$ has a spike, it may take a high degree polynomial to capture the spike and hence require too many parameters to be estimated. Figure 4.1 shows an example of the number of polynomials required to capture a spike in a function. It can be seen that even a polynomial of degree 20 cannot approximate $\theta(t)$ well.

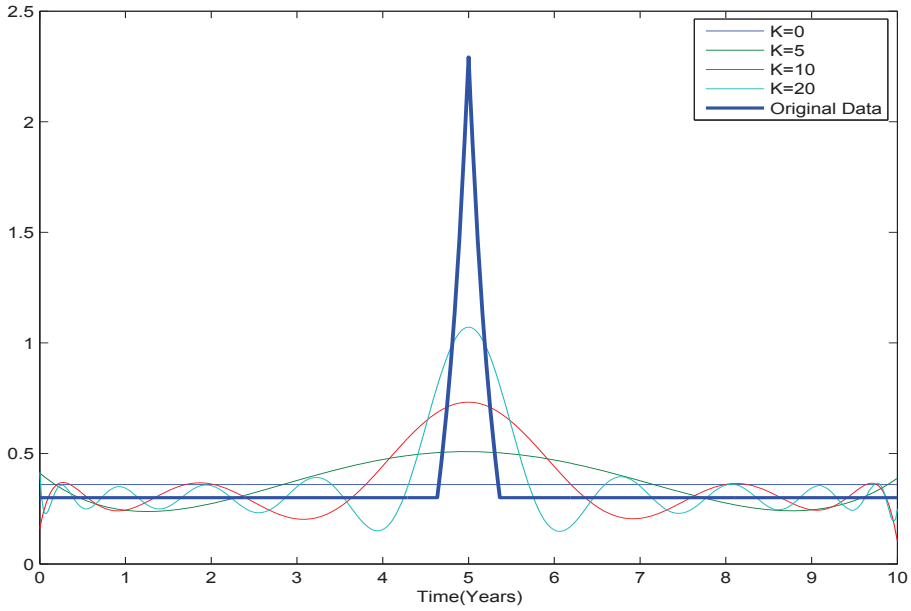


Figure 4.1: Function with a spike and its approximations using Legendre polynomials of degrees $K = 0, 5, 10$ and 20

We apply an estimation approach similar in spirit to the sieve method to determine how fast we can increase the number of parameters so that the consistency of our estimators can be secured. The sieve method is a nonparametric estimation method whose mathematical groundwork was developed by Grenander (1981). In the context of stationary time series, asymptotic consistency and normality have been studied, for example, in Chen (2007) and Bienrens (2011). In the context of continuous diffusion processes, some existing works include those of Nguyen and Pham (1982), Mckeague (1986), Beder (1987), Stone and Huang (2003), Prakasa Rao (2004). Theses authors prove, under different assumptions, the consistency and asymptotic normality of the sieve estimator for a time-dependent drift component when the number of paths of the whole diffusion process increases to infinity. Genon-Catalot et.al (1992) estimate the diffusion term using the sieve method based on discrete data and prove consistency under the assumption that the sampling in-

terval converges to zero. We provide a more complete overview of the literature in Section 1.2.4.

The sieve-type method we are investigating is not the same as the existing ones in the literature such as that of Nguyen and Pham(1982), where the authors consider the asymptotic scheme that the number of independent paths increases to infinity. Our statistical inference is based on one single realization of the random process and the asymptotics is studied as $T \to \infty$, where T is the length of the time interval over which the process is observed. From a practical point of view, it is quite often the case that we have only one single realization available. In this chapter, we show that we can consistently estimate a projection of the drift function onto a finite dimensional space. From a mathematical point of view, there are two technical difficulties in our approach:

1) The usual limit theorems based on i.i.d random variables (processes) are not applicable.

2) We are estimating a projection of the time-dependent function $\{\theta(t)\}$ onto a subspace of $L^2([0,T],\mu(dt))$, with the measure $\mu(dt) \triangleq \frac{1}{\sigma^2(t,r(t))}dt$. Even if the dimensions of projection spaces are the same, the set of parameters we estimate depends on time T, which makes our parametrization different from those used in the existing literature. Therefore, an extension of the sieve method to our case is not trivial.

The layout of the rest of this chapter is as follows:

- Section 4.2 demonstrates that for time series data generated from Brownian motion, the parameters in the drift component cannot be estimated without statistical error, no matter how densely we sample the process.

- Section 4.3 derives a sequence of restricted maximum likelihood estimators of a projection of the drift function onto a finite dimensional space. Exact confidence intervals are derived, and a hypothesis-testing procedure is proposed to determine the dimension of the parameter space. Moreover, the asymptotic consistency and integrated square error of the resulting estimator are also studied.

- Section 4.4 extends the above results to a more general class of models motivated by Nguyen and Pham (1982).

- Section 4.5 presents simulation results, and Section 4.6 draws concluding remarks.

4.2 Discretely Sampled Data

Here we show that for discretely sampled data from (4.1), we cannot perfectly estimate the drift function, no matter how densely we sample the process. This well-known result is in contrast to the estimation of the diffusion parameter, which can be perfectly estimated as the sampling interval shrinks to zero. Because we were unable to find a general theorem stating this finding in the literature, in the following we provide a statement for a special case.

Theorem 4.1. *Suppose we observe data from the following process:*

$$dr(t) = \theta(t)dt + dW(t), \quad 0 \leq t \leq 1, \tag{4.2}$$

where $\theta(t) = \theta_0 + \theta_1 t$. *We denote by* $\vec{\theta} = (\theta_0, \theta_1)$ *the parameters of interest. Assume that we sample the process at discrete time points* $t_i = \frac{i}{n}, i = 1, 2, \cdots, n$. *Then the maximum likelihood estimator* $\hat{\vec{\theta}}_n = (\hat{\theta}_{0_n}, \hat{\theta}_{1_n})$ *does not converge to the true parameter vector* $\vec{\theta}$ *in probability as* $n \to \infty$, *i.e., when we sample the process more densely.*

Proof. It follows from (4.2) and the parameterization of the drift parameter that

$$r(t) = r(0) + \sum_{j=1}^{2} \frac{\theta_{j-1}}{j} t^j + W(t).$$

Then we have

$$r(\frac{i+1}{n}) - r(\frac{i}{n}) \;=\; \sum_{j=1}^{2} \frac{\theta_{j-1}}{j}[(\frac{i+1}{n})^j - (\frac{i}{n})^j] + W(\frac{i+1}{n}) - W(\frac{i}{n}), i = 0, \cdots, n-1. \tag{4.3}$$

Let $Y_{i+1} \triangleq r(\frac{i+1}{n}) - r(\frac{i}{n}), Y \triangleq (Y_1, Y_2, \cdots, Y_n)', \epsilon_{i+1} \triangleq W(\frac{i+1}{n}) - W(\frac{i}{n}), \epsilon \triangleq (\epsilon_1, \cdots, \epsilon_n)'$ and

$$X \triangleq \begin{pmatrix} \frac{1}{n} & \frac{1}{2n^2} \\ \frac{1}{n} & \frac{3}{2n^2} \\ \vdots & \vdots \\ \frac{1}{n} & \frac{2n-1}{2n^2} \end{pmatrix}.$$

Then

$$\epsilon_{i+1} \sim N(0, \frac{1}{n}), \quad Cov(\epsilon_i, \epsilon_j) = 0, \quad i \neq j,$$

and (4.3) can be rewritten as

$$Y = X\vec{\theta} + \epsilon.$$

Using standard methods, it is easy to verify that the maximum likelihood estimator of $\vec{\theta}$ is given by $\hat{\vec{\theta}}_n = (X'X)^{-1}X'Y$, and $\hat{\vec{\theta}}_n \sim MVN(\vec{\theta}, (nX'X)^{-1})$. It follows from elementary algebra that

$$nX'X = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} - \frac{1}{12n^2} \end{pmatrix}.$$

Then

$$\begin{aligned}(nX'X)^{-1} &= \begin{pmatrix} \frac{4n^2-1}{n^2-1} & \frac{12n^2}{2(1-n^2)} \\ \frac{12n^2}{2(1-n^2)} & \frac{12n^2}{n^2-1} \end{pmatrix} \\ &\longrightarrow \begin{pmatrix} 4 & -6 \\ -6 & 12 \end{pmatrix} \; as \; n \to \infty.\end{aligned}$$

This $\vec{\hat{\theta}}_n$ converges in distribution to a vector of normal random variables with mean $\vec{\theta}$ and variance matrix $\begin{pmatrix} 4 & -6 \\ -6 & 12 \end{pmatrix}$, which implies that $\vec{\hat{\theta}}_n$ does not converge in probability to $\vec{\theta}$. □

This result can be recovered by finding the estimation error for the maximum likelihood method based on a continuous trajectory.

4.3 Continuously Sampled Data

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{\mathcal{F}_t\}$ be a filtration on $(\Omega, \mathcal{F})$. Suppose we observe one trajectory of the process (4.1) over $[0, T]$ and we are interested in estimating $\theta(\cdot)$. We make the following assumptions:

- Both $\theta(\cdot)$ and $\sigma(\cdot, \cdot)$ satisfy the linear growth and Lipschitz conditions (e.g., Lipster and Shiryaev (1977), Mao and Yuan (2006)), to ensure the existence of a unique solution to (4.1).

- There exist $\epsilon, M > 0$ such that $\epsilon < \sigma(t, r(t)) < M, \forall t \in [0, T]$. The boundedness of the diffusion coefficient is realistic, because uncertainty in the financial markets should neither vanish nor explode. Moreover, we assume that $\theta(\cdot)$ is bounded on $[0, T]$, which is also a reasonable assumption because the drift function should not explode.

- Only one single continuous realization of $\{r(t) : 0 \leq t \leq T\}$ is observed. In financial and economic problems, we usually have a chance to observe more data of a process, such as the stock index and interest rate throughout time, instead of observing the same process many times. This is our motivation to look at the asymptotic properties of estimators when $T \to \infty$.

- For each t, the diffusion function value $\sigma(t, r(t))$ is known, as otherwise, it can be estimated without error from the quadratic variation of the process.

4.3.1 Description of the Methodology

Before we discuss the proposed estimation method and hypothesis testing for the time-dependent drift function $\theta(t)$, we introduce several notations that will be useful in the rest of this thesis.

For a given realization $\{r(t), t \in [0, T]\}$ of the process (4.1), we define a measure on $[0, T]$ by $\mu(dt) \triangleq \frac{1}{\sigma^2(t,r(t))}dt$ and consider the parameter space $L^2([0, T], \mu(dt))$. The boundedness of $\sigma(\cdot, \cdot)$ and $\theta(\cdot)$ implies that $\theta(\cdot) \in L^2([0, T], \mu(dt))$. We have chosen to use the space $L^2([0, T], \mu(dt))$ in our analysis because it is a Hilbert space. Our focus will be on estimation of a projection of $\{\theta(t) : 0 \leq t \leq T\}$ onto a finite dimensional subspace within $L^2([0, T], \mu(dt))$. In addition, we denote the inner product on $L^2([0, T], \mu(dt))$ by $< \cdot, \cdot >_\mu$; that is, $< f, g >_\mu \triangleq \int_0^T f(t)g(t)\mu(dt)$.

The following notations will be useful in describing our projections. Let $\vec{p}_{T,K}(t) \triangleq (p_{0,T}(t), p_{1,T}(t), \cdots, p_{K,T}(t))', \quad 0 \leq t \leq T$, be a vector of linearly independent functions from $L^2([0, T], \mu(dt))$ and $V_{T,K} \triangleq span\{p_{j,T}(t) : j = 0, \cdots, K, \quad 0 \leq t \leq T\}$. Then $V_{T,K} \subset L^2([0, T], \mu(dt))$. We define $M_{T,K}$ to be the orthogonal projection operator from $L^2([0, T], \mu(dt))$ onto $V_{T,K}$. For any $\{u(t) : 0 \leq t \leq T\} \in L^2([0, T], \mu(dt))$, it follows from basic properties of projections

in Hilbert spaces that

$$M_{T,K}(u)(t) = \vec{p}_{T,K}(t)'\Phi_{T,K}(u), \tag{4.4}$$

where $\Phi_{T,K}(u)$ represents the vector of coefficients of the projection $M_{T,K}(u)$ in the coordinates represented by the basis system $\vec{p}_{T,K}(t)'$. More explicitly, the operator $\Phi_{T,K}(u)$ is defined by $\Phi_{T,K}(u) = \left(\int_0^T \vec{p}_{T,K}(t)\vec{p}_{T,K}(t)'\mu(dt)\right)^{-1} \int_0^T u(t)\vec{p}_{T,K}(t)\mu(dt)$. We present a proof of (4.4) in Appendix 4.7.3.

To obtain useful theoretical results, in this work we choose $p_{j,T}(t)$ to be of a specific form, which is often a transformation of a particular set $q_{j,T}(t), j \geq 0$. Here we briefly describe the construction. Suppose that $\vec{R}(t) = (R_0(t), R_1(t), \cdots, R_i(t), \cdots)$, $-1 \leq t \leq 1$, is an orthonormal basis on $L^2([-1,1], dt)$, such as normalized Legendre polynomials or normalized trigonometric polynomials (Appendix 4.7.2). We define $q_{i,T}(t) \triangleq R_{i,T}(\frac{2t}{T} - 1), 0 \leq t \leq T, i \geq 0$. Then $\vec{q}_T(t) \triangleq \vec{R}(\frac{2t}{T} - 1)$ is an orthogonal basis for $L^2([0,T], dt)$ and $\int_0^T q_{i,T}(t)q_{j,T}(t)dt = \frac{T}{2}\delta_{ij}$, where δ is the Kronecker delta. We emphasize that $q_{i,T}(t), i \geq 0$ are deterministic functions, but $p_{i,T}(t), i \geq 0$ may be chosen to be functions of $r(t)$. We provide a detailed description in Section 4.3.2.

The Methodology

Since $L^2([0,T], \mu(dt))$ is an infinite dimensional space, the maximum likelihood estimation of $\{\theta(t) : 0 \leq t \leq T\}$ over $L^2([0,T], \mu(dt))$ is not directly applicable. For any fixed length T of the process path and fixed K, our objective is to estimate $M_{T,K}(\theta)$, which is the projection of the drift function $\{\theta(t) : 0 \leq t \leq T\}$ onto $V_{T,K}$. More specifically, we provide a point-wise estimator of the function $M_{T,K}(\theta)$ for each $t \in [0,T]$, which we denote by $\hat{M}_{T,K}(\theta)(t)$. We emphasize that $M_{T,K}(\theta)(t)$ changes with both T and K. Therefore, we are not estimating a fixed quantity of interest as we increase T. Figure 4.2 shows an example of the projection of $\theta(t)$ onto $V_{T,K=1}$ over different horizons. The horizons T are taken to be $1, 4$ and 5 years

and $\theta(t) = 10 + 1.2t + 0.13t^2 - 0.05t^3$. It can be seen that even for the same t and K, the projection $M_{T,K}(\theta)(t)$ is different for different T. We also considered projection of θ onto the space of quadratic functions. In this case, the projection was closer to the true function, but the dependence on T was still visible.

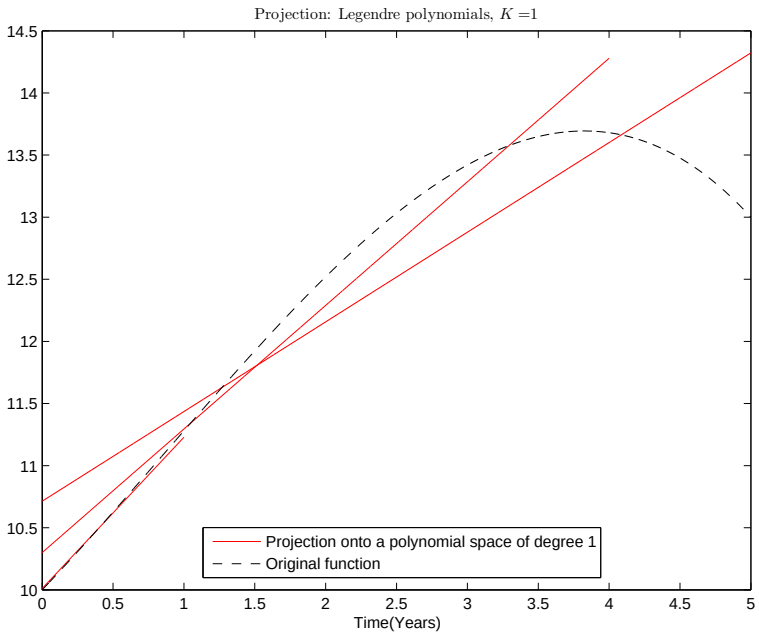


Figure 4.2: Projection of $\theta(t)$ onto the space of linear functions over different time horizons: $T = 1, 4$ and 5 years

It is expected that more parameters, or a projection of the drift function onto a larger space, can be estimated with a pre-determined accuracy as the observed trajectory becomes longer. In other words, we allow the dimension parameter K to be a nondecreasing function of T, denoted by K_T. However, we must control the speed at which K_T increases. A rapidly increasing number of parameters can explode the variance of our estimator for $M_{T,K}(\theta)$. The idea is similar in spirit to that of the sieve method developed by Grenander (1981). When $\sigma(t, r(t)) = 1$ for model (4.1), Grenander has considered the estimation of the unknown $\theta(\cdot)$ based

on n observed trajectories of $\{r_n(t) : 0 \leq t \leq T\}$ and studied the asymptotic properties of a sequence of maximum likelihood estimators of $\theta(\cdot)$ (based on a sieve) when $n \to \infty$. However, as we have emphasized, we consider the asymptotic scheme to be $T \to \infty$. The following list outlines the methods we propose:

P 1. We derive a sequence of restricted maximum likelihood estimators of $M_{T,K}(\theta)$ over the sieve spaces $V_{T,K}$, which is the maximum likelihood estimator for model (4.1) with a parameterized drift function $\theta(t)$ in $V_{T,K}$. Then for each t we derive a confidence interval for $M_{T,K}(\theta)(t)$ based on the distribution of the maximum likelihood estimator $\hat{M}_{T,K_T}(\theta)(t)$.

P 2. For any finite length of data T, it is of practical interest to determine the dimension K of the projection space so that we can estimate $M_{T,K_T}(\theta)$ at a given level of accuracy. To answer this question, a hypothesis-testing procedure is developed to test the null hypothesis:

H_0: the degree of the time-dependent drift function is no larger than K_0.

The idea is to assume that the degree of the drift function is less than a large number K_{max}, and then test whether the coefficients for basis functions of degrees higher than K_0 are zero. For this purpose, we derive chi-square test statistics under the null hypothesis.

P 3. We prove that if K_T increases at a controlled speed with T, the sequence of estimators $\hat{M}_{T,K_T}(\theta)(t)$ for $M_{T,K_T}(\theta)(t)$ is weakly consistent for each t. By "weakly consistent", we mean that $\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)$ converges to 0 in probability; that is, the difference between our estimator and the estimation target converges to zero in probability. The idea is to find conditions such that the variance of $\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)$ vanishes as $T \to \infty$ and then make use of the Chebyshev's inequality to prove that $\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)$ converges to zero in probability.

We would like to emphasize that the focus of the above procedures is on the estimation of $M_{T,K}(\theta)(t)$, instead of the true drift function $\theta(t)$. The reason for this is that for any finite-length trajectory of the process, we can estimate only a finite number of parameters at a given level of accuracy.

4.3.2 The Maximum Likelihood Estimator of the Projected Drift and Its Properties

In the following, we derive the maximum likelihood estimator of $\theta(t)$ over a restricted parameter space $V_{T,K}$. Since $\theta(\cdot) \in L^2([0,T], \mu(dt))$, we can write

$$\theta(t) = M_{T,K}(\theta)(t) + M^{\perp}_{T,K}(\theta)(t) = \vec{p}_{T,K}(t)'\Phi_{T,K}(\theta) + M^{\perp}_{T,K}(\theta)(t), \quad 0 \le t \le T,$$

where the operator $M^{\perp}_{T,K}$ is defined to be $I - M_{T,K}$, with I being the identity operator. Since $p_{i,T}(t), i \ge 0$ are linearly independent functions on $[0,T]$, we have a unique representation of $\Phi_{T,K}(\theta) = (\theta_{0,T}, \theta_{1,T}, \cdots, \theta_{K,T})'$. Therefore, we have the following representation:

$$\theta(t) = \sum_{j=0}^{K} \theta_{j,T} p_{j,T}(t) + M^{\perp}_{T,K}(\theta)(t), \quad 0 \le t \le T.$$

Let P_r^T be the probability measure generated by the process $\{r(t) : 0 \le t \le T\}$ on the space $(C[0,T], \mathcal{B}_T)$, where $C[0,T]$ denotes the space of the continuous functions endowed with the supremum norm and $\mathcal{B}_T$ corresponding Borel σ-algebra. Let P_η^T be the probability measure induced by the strong solution to the equation $d\eta(t) = \sigma(t, \eta(t))dW(t), \quad 0 \le t \le T$. The boundedness of $\theta(\cdot)$ and $\sigma(\cdot,\cdot)$ imply that

$$P(\omega \in \Omega : |\frac{\theta(t)}{\sigma(t, r(t))}| < \infty) = 1, \quad \forall\ 0 \le t \le T, \tag{4.5}$$

$$P(\omega \in \Omega : \int_0^T \frac{\theta(t)^2}{\sigma^2(t, r(t))} dt < \infty) = 1. \tag{4.6}$$

Then we have $P_r^T \ll P_\eta^T$ (Lipster & Shiryaev (1974)) and

$$\begin{aligned}
&\frac{dP_r^T}{dP_\eta^T} \\
= \;& \exp\{\int_0^T \frac{\sum_{j=0}^K \theta_{j,T} p_{j,T}(t) + M_{T,K}^{\perp}(\theta)(t)}{\sigma^2(t,r(t))} dr(t) - \frac{1}{2}\int_0^T \frac{(\sum_{j=0}^K \theta_{j,T} p_{j,T}(t) + M_{T,K}^{\perp}(\theta)(t))^2}{\sigma^2(t,r(t))} dt \\
= \;& \exp\{\int_0^T \frac{\sum_{j=0}^K \theta_{j,T} p_{j,T}(t) + M_{T,K}^{\perp}(\theta)(t)}{\sigma^2(t,r(t))} dr(t) \\
& -\frac{1}{2}\int_0^T \frac{(\sum_{j=0}^K \theta_{j,T} p_{j,T}(t))^2 + (M_{T,K}^{\perp}(\theta)(t))^2}{\sigma^2(t,r(t))} dt\},
\end{aligned}$$

where in the second equation we used the fact that $M_{T,K}^{\perp}(\theta)$ is orthogonal to $p_{j,T}, j \in \{0, \cdots, K\}$ in $L^2([0,T], \mu(dt))$. By treating $M_{T,K}^{\perp}(\theta)(t)$ as a nuisance parameter, we can derive the maximum likelihood estimator of $\Phi_{T,K}(\theta) = (\theta_{0,T}, \theta_{1,T}, \cdots, \theta_{K,T})'$ by maximizing the log-likelihood function, which is given by

$$\begin{aligned}
& l(\Phi_{T,K}(\theta)) \\
= \;& \int_0^T \frac{\sum_{j=0}^K \theta_{j,T} p_{j,T}(t) + M_{T,K}^{\perp}(\theta)(t)}{\sigma^2(t,r(t))} dr(t) - \frac{1}{2}\int_0^T \frac{(\sum_{j=0}^K \theta_{j,T} p_{j,T}(t))^2 + (M_{T,K}^{\perp}(\theta)(t))^2}{\sigma^2(t,r(t))} dt \\
= \;& \int_0^T \frac{\vec{p}_{T,K}(t)'\Phi_{T,K}(\theta) + (M_{T,K}^{\perp}(\theta)(t))}{\sigma(t,r(t))^2} dr(t) - \frac{1}{2}\int_0^T \frac{(\vec{p}_{T,K}(t)'\Phi_{T,K}(\theta))^2 + (M_{T,K}^{\perp}(\theta)(t))^2}{\sigma(t,r(t))^2} dt.
\end{aligned}$$

By differentiating with respect to $\theta_{m,T}, m \in \{0, \cdots, K\}$ we can find that the score function is given by

$$\frac{\partial l(\Phi_{T,K}(\theta))}{\partial \theta_{m,T}} = \int_0^T \frac{p_{m,T}(t)}{\sigma(t,r(t))^2} dr(t) - \int_0^T \frac{\vec{p}_{T,K}(t)'\Phi_{T,K}(\theta) p_{m,T}}{\sigma(t,r(t))^2} dt, \quad \forall\ 0 \le m \le K.$$

Therefore, the maximum likelihood estimator $\hat{\Phi}_{T,K}(\theta)$ of $\Phi_{T,K}(\theta)$ satisfies

$$\int_0^T \frac{p_{m,T}(t)}{\sigma(t,r(t))^2} dr(t) = \int_0^T \frac{p_{m,T}(t)[\vec{p}_{T,K}(t)'\hat{\Phi}_{T,K}(\theta)]}{\sigma(t,r(t))^2} dt, \quad 0 \le m \le K,$$

or using vector notation,

$$\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2} dr(t) = \int_0^T \frac{\vec{p}_T(t)[\vec{p}_{T,K}(t)'\hat{\Phi}_{T,K}(\theta)]}{\sigma(t,r(t))^2} dt = A_{T,K}\hat{\Phi}_{T,K}(\theta),$$

where $A_{T,K} \triangleq \int_0^T \frac{\vec{p}_T(t)\vec{p}_{T,K}(t)'}{\sigma(t,r(t))^2}dt$.

By solving the last equation, we obtain the following maximum likelihood estimator of $\Phi_{T,K}(\theta)$:

$$\hat{\Phi}_{T,K}(\theta) = A_{T,K}^{-1} \int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2} dr(t), \tag{4.7}$$

where the invertibility of $A_{T,K}$ is proved in Appendix 4.7.4. From (4.7), now we can find the maximum likelihood estimator of $M_{T,K}(\theta)(t)$ for any fixed t:

$$\hat{M}_{T,K}(\theta)(t) \triangleq \vec{p}_{T,K}(t)'\hat{\Phi}_{T,K}(\theta) = \vec{p}_{T,K}(t)' A_{T,K}^{-1} \int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2} dr(t). \tag{4.8}$$

To derive explicit distributions of the maximum likelihood estimator $\hat{\Phi}_{T,K}(\theta)$ and $\hat{M}_{T,K}(\theta)(t)$, in the following we replace $dr(t)$ in both (4.7) and (4.8) by the right-hand side of equation (4.1):

$$\begin{aligned}
\hat{\Phi}_{T,K}(\theta) &= A_{T,K}^{-1} \int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2} [\theta(t)dt + \sigma(t,r(t))dW(t)] \\
&= A_{T,K}^{-1} \int_0^T \frac{\vec{p}_{T,K}(t)\theta(t)}{\sigma(t,r(t))^2} dt + A_{T,K}^{-1} \int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))} dW(t) \\
&= \Phi_{T,K}(\theta) + A_{T,K}^{-1} \int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))} dW(t).
\end{aligned} \tag{4.9}$$

Using this result, we have the following representation of the MLE of the projection $M_{T,K}(\theta)$ at time t:

$$\begin{aligned}
\hat{M}_{T,K}(\theta)(t) &= \vec{p}_{T,K}(t)'[\Phi_{T,K}(\theta) + A_{T,K}^{-1} \int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))} dW(t)], \\
&= M_{T,K}(\theta)(t) + \vec{p}_{T,K}(t)' A_{T,K}^{-1} \int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))} dW(t).
\end{aligned} \tag{4.10}$$

From (4.9) and (4.10), the estimators $\hat{\Phi}_{T,K}(\theta)$ and $\hat{M}_{T,K}(\theta)(t)$ have been expressed as sums of the parameter of interest and error terms. Note that $\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))} dW(t)$ is common to both terms.

To make our analysis feasible, we choose $\vec{p}_{T,K}(t) = \sigma(t, r(t))\vec{q}_{T,K}(t)$ so that the distribution of $\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))} dW(t)$ is in an explicit closed form. Since by definition $\vec{q}_{T,K}(t)$ are orthogonal functions in $L^2([0,T], dt)$, we have

$$\begin{aligned} A_{T,K} &= \int_0^T \frac{\vec{p}_{T,K}(t)'\vec{p}_{T,K}(t)}{\sigma^2(t, r(t))} dt \\ &= \int_0^T \vec{q}_{T,K}(t)'\vec{q}_{T,K}(t) dt = \frac{T}{2} I_{(K+1)\times(K+1)}. \end{aligned}$$

It is then easy to verify that

$$\begin{aligned} \int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t, r(t))} dW(t) = \int_0^T \vec{q}_{T,K}(t) dW(t) &\sim MVN(0, \int_0^T \vec{q}_{T,K}(t)\vec{q}_{T,K}(t)' dt) \\ &= MVN(0, \frac{T}{2} I_{(K+1)\times(K+1)}). \end{aligned}$$

The representation (4.9) implies that

$$\hat{\Phi}_{T,K}(\theta) - \Phi_{T,K}(\theta) \sim MVN(0, \frac{2}{T} I_{(K+1)\times(K+1)}), \tag{4.11}$$

and the representation (4.10) implies that

$$\begin{aligned} \frac{\hat{M}_{T,K}(\theta)(t) - M_{T,K}(\theta)(t)}{\sigma(t, r(t))} &= \vec{q}_{T,K}(t)' A_{T,K}^{-1} \int_0^T \vec{q}_{T,K}(t) dW(t) \\ &= \frac{2}{T} \vec{q}_{T,K}(t)' \int_0^T \vec{q}_{T,K}(t) dW(t) \\ &\sim N(0, \frac{2}{T} \vec{q}_{T,K}(t)'\vec{q}_{T,K}(t)). \end{aligned} \tag{4.12}$$

The forms (4.11) and (4.12) are important results for this chapter, as they will be used in developing theoretical results throughout later sections. The definition of the basis function we use is convenient, since then $A_{T,K} = \frac{T}{2} I_{(K+1)\times(K+1)}$ and hence, from (4.11) and (4.12), the variances of $\hat{\Phi}_{T,K}(\theta)$ and $\hat{M}_{T,K}(\theta)(t)$ are proportional to $\frac{1}{T}$.

4.3.3 Asymptotic Consistency: a Sieve-type Approach

In this section, we allow the dimension parameter K to increase with T, which we emphasize by using K_T, and prove that the estimator $\hat{M}_{T,K_T}(\theta)(t)$ converges in probability to $M_{T,K_T}(\theta)(t)$ for any fixed t. Under an additional condition that $M_{T,K_T}(\theta)(t)$ converges in probability to $\theta(t)$ at least as fast as the speed at which $\hat{M}_{T,K_T}(\theta)(t)$ converges to $M_{T,K_T}(\theta)(t)$, we can also prove that $\hat{M}_{T,K_T}(\theta)(t)$ converges in probability to the true drift $\theta(t)$. We discuss this additional condition at the end of this section.

In the following, we prove that $\hat{M}_{T,K_T}(\theta)(t)$ is weakly consistent for estimating $M_{T,K_T}(\theta)(t)$, provided that K_T does not grow too fast with T. The proof of the theorem is presented in Appendix 4.7.5.

Theorem 4.2. *Let $\{\vec{R}(t) : -1 \le t \le 1\}$ be either normalized Legendre polynomials or trigonometric polynomials (Appendix 4.7.2). The following results hold:*

(i) If $K_T = K$, i.e., K_T does not increase with T, then for each fixed $t \ge 0$,

$$\sqrt{T}\left(\frac{\hat{M}_{T,K}(\theta)(t) - M_{T,K}(\theta)(t)}{\sigma(t, r(t))}\right) \to N(0, 2\sum_{i=0}^{K} R_i^2(-1)) \text{ in distribution, as } T \to \infty.$$

In other words, $\hat{M}_{T,K}(\theta)(t)$ is a weakly consistent estimator of $M_{T,K}(\theta)(t)$, and the convergence rate of $\hat{M}_{T,K}(\theta)(t)$ is at least as fast as $\frac{1}{\sqrt{T}}$ regardless of the choice of $\vec{R}(t)$.

(ii) If K_T changes with T, then $\hat{M}_{T,K_T}(\theta)(t)$ is a weakly consistent estimator of $M_{T,K_T}(\theta)(t)$ under the following condition:

(C1) $\lim_{T\to\infty} \frac{K_T^2}{T} = 0$ if $\{\vec{R}(t) : -1 \le t \le 1\}$ are the normalized Legendre polynomials.

(C2) $\lim_{T\to\infty} \frac{K_T}{T} = 0$ if $\{\vec{R}(t) : -1 \le t \le 1\}$ are the normalized trigonometric polynomials.

So far we have emphasized that we estimate a projection of the drift function onto a suitable subspace, instead of the drift function itself. The following result states that, under additional conditions, $\hat{M}_{T,K_T}(\theta)(t)$ converges in probability to the true drift function $\theta(t)$ as $T \to \infty$. Loosely speaking, when we have a subspace large enough to approximate the drift function well and the trajectory length is long enough to accurately estimate all parameters determining the subspace, $\hat{M}_{T,K_T}(\theta)(t)$ is close to the true drift function $\theta(t)$.

Theorem 4.3. *Assume that condition* $\mathcal{C}1$ *or* $\mathcal{C}2$ *(depending on the choice of basis system* $\vec{R}(t)$*) holds. Then for every* $t \in [0, T]$*,* $\hat{M}_{T,K_T}(\theta)(t)$ *converges in probability to the drift function* $\theta(t)$ *if the following condition is also satisfied:*

($\mathcal{C}3$*) For any* $\epsilon, \eta > 0$*, there exists* $T_{\epsilon,\eta}$ *such that for all* $T > T_{\epsilon,\eta}$*,*

$$P(|M^{\perp}_{T,K_T}(\theta)(t)| > \epsilon) < \eta.$$

Proof. In fact, under condition $\mathcal{C}3$, the arbitrariness of ϵ and η implies that

$$\lim_{T\to\infty} |\theta(t) - M_{T,K_T}(\theta)(t)| = 0 \text{ in probability.} \tag{4.13}$$

On the other hand, it follows from Theorem 4.2 that under condition ($\mathcal{C}1$) or ($\mathcal{C}2$), $\hat{M}_{T,K_T}(\theta)(t)$ is a weakly consistent estimator of $M_{T,K_T}(\theta)(t)$. Using the definition of convergence in probability or Slutsky's theorem, this result in conjunction with (4.13) implies that $\hat{M}_{T,K_T}(\theta)(t)$ converges in probability to the drift function $\theta(t)$. □

It is well known that the larger the projection space is, the smaller the residue part is. Condition $\mathcal{C}3$ is clearly satisfied if $\{\theta(t) : 0 \le t \le T\} \in V_{T,K^*}$ for some finite K^*. For example, if $\{\theta(t) : t \ge 0\}$ is a polynomial of degree 10 and the Legendre polynomial basis system with $K \ge 10$ is chosen as the subspace $V_{T,K}$ for estimation purposes, then $\hat{M}_{T,K}(\theta)(t)$ is a weakly consistent estimator of the true drift function $\theta(t)$. We believe that condition $\mathcal{C}3$ can be satisfied by a more general

class of functions, e.g., for functions that are continuous and converge to a constant as $t \to \infty$. Unfortunately, we do not have a proof for this conjecture.

4.3.4 Hypothesis Testing of the Dimension of the Parameter Space

To apply the estimation methods proposed in Sections 4.3.2 and 4.3.5, we need to specify $V_{T,K}$. Assume that we have chosen a basis system $\{q_{0,T}(t), q_{1,T}(t), \cdots\}$ such as the Lengendre polynomials or trigonometric polynomials. The next question is how to decide on the number of basis functions that should be applied to fit data of a given length T; i.e., *what K should be used?* As we will see in Section 4.3.6, a high number of basis functions can inflate the variance and lead to a less accurate estimator of the projection $M_{T,K}(\theta)(t)$. In this section, we introduce a hypothesis-testing procedure to determine the dimension of the parameter space. For this, we consider spaces only with dimensions up to a prescribed number K_{max}. Thus, for a given number K_0, we are interested in testing the following null hypothesis

H_0 : *the coefficients corresponding to basis functions of degrees higher than K_0 but smaller than K_{max} are zero .*

In other words, under H_0, the degree of the time-dependent drift functions is either larger than K_{max} or no larger than K_0. If the null hypothesis is true, then we prefer the more parsimonious parameter space V_{T,K_0} rather than the larger parameter space $V_{T,K_{max}}$.

Let us define a matrix

$$H = (H_1 \quad H_2),$$

where $H_1 = 0_{(K_{max}-K_0)\times(K_0+1)}$ is has elements zero, and $H_2 = I_{(K_{max}-K_0)\times(K_{max}-K_0)}$ is an identity matrix. From (4.11), we know that $\hat{\Phi}_{T,K_{max}}(\theta) - \Phi_{T,K_{max}}(\theta) \sim$

$MVN(0, A^{-1}_{T,K_{max}})$. Therefore,

$$H(\hat{\Phi}_{T,K_{max}}(\theta) - \Phi_{T,K_{max}}(\theta)) \sim MVN(0, HA^{-1}_{T,K_{max}}H'),$$

or more explicitly,

$$(\hat{\Phi}^{K_{max}}_{T,K_0}(\theta) - \Phi^{K_{max}}_{T,K_0}(\theta)) \sim MVN(\vec{0}, HA^{-1}_{T,K_{max}}H'),$$

where

$$\Phi^{K_{max}}_{T,K_0}(\theta) \triangleq H\Phi_{T,K_{max}}(\theta) = (\theta_{K_0+1}, \theta_{K_0+2}, \cdots, \theta_{K_{max}})',$$

and

$$\hat{\Phi}^{K_{max}}_{T,K_0}(\theta) \triangleq H\hat{\Phi}_{T,K_{max}}(\theta) = (\hat{\theta}_{K_0+1,T}, \hat{\theta}_{K_0+2,T}, \cdots, \hat{\theta}_{K_{max},T}).$$

The purpose of the matrix H is to reduce the number of parameters from $K_{max}+1$ to $K_{max} - K_0$. Under our null hypothesis, $\Phi^{K_{max}}_{T,K_0}(\theta) = \vec{0}$ and hence

$$\hat{\Phi}^{K_{max}}_{T,K_0}(\theta) \sim MVN(\vec{0}, HA^{-1}_{T,K_{max}}H'),$$

or equivalently,

$$H\hat{\Phi}_{T,K_{max}}(\theta) \sim MVN(\vec{0}, HA^{-1}_{T,K_{max}}H'). \tag{4.14}$$

If we substitute (4.7) into (4.14), we get

$$HA^{-1}_{T,K_{max}} \int_0^T \frac{\vec{p}_{T,K_{max}}(t)}{\sigma(t, r(t))^2} dr(t) \sim MVN(\vec{0}, HA^{-1}_{T,K_{max}}H'). \tag{4.15}$$

It follows from the proof in Appendix 4.7.4 that $HA^{-1}_{T,K_{max}}H'$ is a positive definite matrix. Therefore, a square root of the matrix $HA^{-1}_{T,K_{max}}H'$ exists, and the representation (4.15) implies

$$[HA^{-1}_{T,K_{max}}H']^{-1/2} HA^{-1}_{T,K_{max}} [\int_0^T \frac{\vec{p}_{T,K_{max}}(t)}{\sigma(t, r(t))^2} dr(t)] \sim MVN(\vec{0}, I_{(K_{max}-K_0)\times(K_{max}-K_0)}).$$

Now we can derive a chi-square test statistic to test H_0. Under the null hypothesis, we have

$$\begin{aligned} & [\int_0^T \frac{\vec{p}_{T,K_{max}}(t)}{\sigma(t,r(t))^2} dr(t)] A_{T,K_{max}}^{-1} H' [H A_{T,K_{max}}^{-1} H']^{-1} H A_{T,K_{max}}^{-1} [\int_0^T \frac{\vec{p}_{T,K_{max}}(t)}{\sigma(t,r(t))^2} dr(t)] \\ \sim \; & \chi^2(K_{max} - K_0), \end{aligned} \tag{4.16}$$

where $\chi^2(K_{max}-K_0)$ represents a random variable with χ^2 distribution with $K_{max}-K_0$ degrees of freedom. We illustrate applications of our proposed hypothesis-testing procedure in Section 4.5.

We propose to define the dimension of the parameter space as the smallest K_0 such that H_0 is accepted. With this K_0, we can apply our estimation method described in Section 4.3.2 and Section 4.3.5 to obtain an estimator for $M_{T,K}(\theta)(t)$ and its confidence interval for each $0 \le t \le T$.

4.3.5 Confidence Interval for the Projected Drift

In this section, we derive a confidence interval for the projection of the drift onto $V_{T,K}$ at a fixed t. Since the basis function $\vec{p}_{T,K}(t)$ is allowed to depend on $r(t)$ (e.g. $\vec{p}_{T,K}(t) = \sigma(t,r(t))\vec{q}_{T,K}(t)$ in Section 4.3.2), $M_{T,K}(\theta)(t)$ may also depend on $r(t)$ and hence it can be a random process. In this case, the confidence interval (CI_{lower},CI_{upper}) for $M_{T,K}(\theta)(t)$ has the following meaning:

$$Prob(CI_{lower} \le M_{T,K}(\theta)(t) \le CI_{upper}) = 95\%.$$

Let $z_{0.975}$ and $z_{0.025}$ be the 97.5th and 2.5th quantiles of a standard normal random variable. It follows from (4.12) that $Var\left(\frac{\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)}{\sigma(t,r(t))}\right) = \frac{2}{T}\sum_{i=0}^{K_T} q_{i,T}^2(t)$ and

$$P\left(z_{0.025}\sqrt{\frac{2}{T}\vec{q}_{T,K}(t)'\vec{q}_{T,K}(t)} \le \frac{\hat{M}_{T,K}(\theta)(t) - M_{T,K}(\theta)(t)}{\sigma(t,r(t))} \le z_{0.975}\sqrt{\frac{2}{T}\vec{q}_{T,K}(t)'\vec{q}_{T,K}(t)}\right) = 95\%.$$

Therefore, for each $t \in [0, T]$, a 95% confidence interval for $M_{T,K}(\theta)(t)$ is

$$\left(\hat{M}_{T,K}(\theta)(t) - z_{0.975}\sigma(t, r(t))\sqrt{\frac{2}{T}\vec{q}_{T,K}(t)'\vec{q}_{T,K}(t)}\ , \right.$$
$$\left. \hat{M}_{T,K}(\theta)(t) - z_{0.025}\sigma(t, r(t))\sqrt{\frac{2}{T}\vec{q}_{T,K}(t)'\vec{q}_{T,K}(t)} \right),$$

or equivalently, due to $\vec{p}_{T,K}(t) = \sigma(t, r(t))\vec{q}_{T,K}(t)$,

$$\left(\hat{M}_{T,K_T}(\theta)(t) - z_{0.975}\sqrt{\frac{2}{T}\sum_{i=0}^{K_T} p_{i,T}^2(t)}, \hat{M}_{T,K_T}(\theta)(t) - z_{0.025}\sqrt{\frac{2}{T}\sum_{i=0}^{K_T} p_{i,T}^2(t)} \right) \quad (4.17)$$

We can measure the size of the confidence interval by its radius, which is given by $\sigma(t, r(t))\sqrt{\frac{2}{T}\sum_{i=0}^{K_T} q_{i,T}^2(t)}$. In combination with (4.30) for the Legendre polynomial case (in the proof of Theorem 4.2 in Appendix 4.7.5), an upper bound for the radius of the 95% confidence interval is $\frac{\sigma(t,r(t))(K_T+1)}{\sqrt{T}}$. Therefore, if we want the estimation error of $\hat{M}_{T,K_T}(\theta)(t)$ to be less than $\alpha\%$ of the true drift function $\theta(t)$, then the requirement on the length of observable data is $T > \left(\frac{(K_T+1)\textit{diffusion size}}{\alpha\%\ \textit{drift size}} \right)^2$, where $\textit{diffusion size} = \sigma(t, r(t))$ and $\textit{drift size} = \theta(t)$. This observation is important because it answers the question *"How much data is needed for a reliable estimate of the parameter of interest ?"* Here the accuracy of our estimation is measured by the relative error $\alpha\%$, and the amount of data is measured by time length T. From our results, it can be seen that there is a tradeoff between the length of data required and the"drift-to-diffusion ratio". In our results, the drift-to-diffusion ratio is measured by $\frac{\textit{drift size}}{\textit{diffusion size}}$. When this ratio is low, we need longer data to achieve our estimation accuracy target.

4.3.6 The Integrated Mean Square Error

In this section, we shift our focus from estimating the projection of the drift to estimating the true drift function. In the following, we still use the same estimator $\hat{M}_{T,K}(\theta)(t)$ as derived in Section 4.3.2, but we treat it as an estimator of $\theta(t)$ instead of $M_{T,K}(\theta)(t)$. The objective is to propose a mechanism for determining the optimal number of parameters to be included for a given length, T, of data. In comparison with the hypothesis testing approach to determine K_T in Section 4.3.4, the criterion in this section is to minimize the integrated mean square error (IMSE) for estimating the true drift function, an error that is a sum of the integrated variance (IVAR) and the integrated square of bias (ISB). An estimator corresponding to a larger K, i.e., with more basis functions and hence more parameters, will have smaller bias because $V_{T,K}$ approximates $L^2([0,T],\mu(dt))$ better; however, the increased variance of the estimator will offset the reduced bias. On the other hand, an estimator based on a smaller K, i.e., with fewer basis functions and hence fewer parameters, will have smaller variance, however the increased bias will offset the reduced variance.

It follows from (4.12) that

$$\frac{\hat{M}_{T,K}(\theta)(t) - M_{T,K}(\theta)(t)}{\sigma(t,r(t))} \sim N(0, \frac{2}{T}\vec{q}_{T,K}(t)'\vec{q}_{T,K}(t)). \tag{4.18}$$

Since the diffusion term $\sigma(t,r(t))$ is allowed to depend on the process itself and hence is a random process, we do not have an explicit distribution of $\hat{M}_{T,K}(\theta)(t)$. To obtain an explicit distribution of $\hat{M}_{T,K}(\theta)(t)$ and calculate the integrated mean square error of $\hat{M}_{T,K}(\theta)(t), t \in [0,T]$, we concentrate in this section only on the special case when $\sigma(t,r(t)) = \sigma(t)$; that is, the diffusion term is modeled as a function of time t only. Under this assumption, we can now multiply both sides of (4.18) by $\sigma(t)$ to obtain

$$\hat{M}_{T,K}(\theta)(t) - M_{T,K}(\theta)(t) \sim N(0, \frac{2}{T}\vec{p}_{T,K}(t)'\vec{p}_{T,K}(t)), \tag{4.19}$$

where $\vec{p}_{T,K}(t) = \sigma(t)\vec{q}_{T,K}(t)$. Another consequence of the assumption $\sigma(t, r(t)) = \sigma(t)$ is that $\{p_{0,T}(t), p_{1,T}(t), \cdots\}$ do not depend on the realization of the process, and neither does $M_{T,K}(\theta)(t)$. Therefore, we can calculate the expectation of $\hat{M}_{T,K}(\theta)(t)$, which is $E[\hat{M}_{T,K}(\theta)(t)] = M_{T,K}(\theta)(t)$. We emphasize that the integration in this section is defined with respect to the measure $\mu(dt) = \frac{1}{\sigma(t)^2}dt$, which again does not depend on the realization of the process.

To formulate the results, we first observe that by (4.19), the IVAR of $\{\hat{M}_{T,K_T}(\theta)(t) : 0 \le t \le T\}$ can be represented as follows:

$$\int_0^T Var[\hat{M}_{T,K}(\theta)(t)]\mu(dt) = \frac{2}{T}\int_0^T \vec{p}_{T,K}(t)'\vec{p}_{T,K}(t)\frac{1}{\sigma(t)^2}dt = K + 1,$$

which is the number of basis functions that determines the subspace $V_{T,K}$. Since $E[\hat{M}_{T,K}(\theta)(t)] = M_{T,K}(\theta)(t)$ and $\{p_{j,T}, j \ge 0\}$ form an orthogonal basis, the ISB for $\hat{M}_{T,K}(\theta)$ is

$$\int_0^T (\theta(t) - M_{T,K}(\theta(t))^2\mu(dt) \;\; = \;\; \sum_{j=K+1}^{\infty} \theta_{j,T}^2 ||p_{j,T}(\cdot)||^2 = \frac{T}{2}\sum_{j=K+1}^{\infty}\theta_{j,T}^2.$$

Therefore, the IMSE of $\{\hat{M}_{T,K}(\theta)(t) : 0 \le t \le T\}$ can be written as

$$IMSE(\hat{M}_{T,K}(\theta)) = K + 1 + \frac{T}{2}\sum_{j=K+1}^{\infty}\theta_{j,T}^2.$$

Since $p_{i,T}, i \ge 0$ is an orthogonal basis, we have

$$\int_0^T \theta(t)^2\mu(dt) = \int_0^T (\sum_{j=0}^{\infty}\theta_{j,T}p_{j,T}(t))^2\mu(dt) = \int_0^T \sum_{j=0}^{\infty}\theta_{j,T}^2(p_{j,T}(t))^2\mu(dt) = \frac{T}{2}\sum_{j=0}^{\infty}\theta_{j,T}^2. \tag{4.20}$$

Therefore, the IMSE of $\{\hat{M}_{T,K}(\theta)(t) : 0 \le t \le T\}$ can be rewritten as

$$IMSE(\hat{M}_{T,K}(\theta)) = K + 1 + \int_0^T \theta^2(t)\mu(dt) - \frac{T}{2}\sum_{j=0}^{K}\theta_{j,T}^2. \tag{4.21}$$

It can be seen from the above formula that there is a tradeoff between the integrated variance and the integrated square of bias of the estimator $\hat{M}_{T,K}(\theta)$. When we increase K, more parameters will be included and the bias of our estimator $\hat{M}_{T,K}(\theta)$ will decrease, but at the same time, the integrated variance will increase. For a given T, we propose a procedure for finding the optimal cut-off, K_{opt}, that minimizes the integrated mean square error. The procedure assumes that $\{\theta_{j,T}, j = 0, 1, \cdots\}$ are known so that a theoretical optimality can be proven. In practice, we work with $\hat{\theta}_{j,T}$ instead of $\theta_{j,T}$ because $\theta_{j,T}$ is unknown.

Step 1: Find a maximal cut-off value K_{max} that has the property

$$\frac{T}{2}\theta_{j,T}^2 \leq 1, \quad \forall j \geq K_{max},$$

i.e., $|\theta_{j,T}| \leq \sqrt{\frac{2}{T}}$ for every $j \geq K_{max}$. The value K_{max} exists due to the completeness of the orthogonal system $\{p_{0,T}(t), p_{1,T}(t), \cdots\}$, which implies that $\int_0^T \theta^2(t)\mu(dt) = \frac{T}{2}\sum_{j=0}^{\infty}\theta_{j,T}^2$ from Parseval's Identity.

Step 2: For $j \in \{0, \cdots, K_{max}\}$, order $\theta_{j,T}^2$ in descending magnitudes. Denote the jth term in this new sequence as $\theta_{O_j,T}^2$, where O_j is a nonnegative integer indicating the index of the same term in the original orthogonal basis system. We allow $j = 0$.

Step 3: Define the optimal cut-off value K_{opt} as the smallest integer I such that $\theta_{O_j,T}^2 \leq \frac{2}{T}, \quad \forall j > I$.

For K_{opt} selected according to the above procedure, we define an estimator of $\theta(\cdot)$ on $[0, T]$ by

$$\hat{\theta}_{T,K_{opt}}(\cdot) \triangleq \sum_{j=0}^{K_{opt}} \hat{\theta}_{O_j,T}\ p_{O_j,T}(\cdot).$$

We now present a property of the proposed estimator. The proof of the theorem is provided in Appendix 4.7.5.

Theorem 4.4. *The estimator* $\hat{\theta}_{T,K_{opt}}(\cdot) = \sum_{j=0}^{K_{opt}} \hat{\theta}_{O_j,T} p_{O_j,T}(\cdot)$ *has the smallest IMSE among all estimators of the form* $\hat{\theta}_F(\cdot) \triangleq \sum_{j\in F} \hat{\theta}_{j,T} p_{j,T}(\cdot)$*, where* F *is an arbitrary subset of nonnegative integers.*

In this section, we have investigated the accuracy of our estimator $\hat{\theta}_{T,K_{opt}}$ in terms of IMSE. In contrast to the previous sections, the parameter of interest is the drift function $\{\theta(t) : 0 \le t \le T\}$, instead of a projection of the drift function onto a suitable subspace. The optimal cut-off value K_{opt} is determined based on the true values of $\{\theta_{j,T}\}$, which are not known a priori. In practice, K_{max} has to be prescribed and $\theta_{j,T}$ has to be estimated from the observed data. Hence we have only an estimator $\hat{K}_{opt}$ of the true K_{opt}.

4.4 Extension to a More General Class of Time-Dependent SDEs

In this section, we look at a more general class of time-dependent diffusion models that includes Brownian motion as a special case. Our objective is still to estimate a projection of the time-dependent component of the drift onto a finite dimensional subspace, given one realization of $r(t)$ on the interval $[0, T]$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{\mathcal{F}_t\}$ be a filtration on $(\Omega, \mathcal{F})$. We assume that the following SDE

$$dr(t) = \theta(t)h(t, r(t))dt + \sigma(t, r(t))dW(t), \quad 0 \le t \le T, \tag{4.22}$$

admits a unique strong solution, which is guaranteed, for example, by the conditions that both $\theta(\cdot)h(\cdot)$ and $\sigma(\cdot, \cdot)$ satisfy Lipschitz and linear growth conditions. The parameter of interest is $\theta(\cdot)$, and we assume that both $h(\cdot)$ and $\sigma(\cdot)$ are known.

One example of the process (4.22) is the following model:

$$dr(t) = \theta(t)r(t)dt + \sigma dW(t), \tag{4.23}$$

which has been studied by Nguyen and Pham (1982), under the assumption that a continuous realization of (4.23) on $[0,T]$ is available and σ is a known constant. Nguyen and Pham apply the sieve method using an increasing sequence of finite dimensional subspaces in a particular Hilbert space to approximate $\theta(\cdot)$. The authors prove consistency and asymptotic normality of the sequence of restricted maximum likelihood estimators when the number of independent realizations of the process tends to infinity. In contrast, we shall study properties of the maximum likelihood estimator when the length of the observation interval T approaches infinity.

Another example of the process (4.22) is a generalized version of Geometric Brownian motion (GBM), where $h(t,x) = x$, and $\sigma(t,x) = \sigma(t)x$. Then

$$\frac{dr(t)}{r(t)} = \theta(t)dt + \sigma(t)dW(t), \quad 0 \le t \le T. \tag{4.24}$$

The setting in this section is similar to the one in Section 4.3. We assume also that $0 < \epsilon < \sigma(t, r(t)) < M$ for some ϵ and M. Moreover, we suppose that $\theta(t)$ is bounded on $[0,T]$ and that $h(\cdot,\cdot)$ is a continuous function on $[0,T] \times \mathbb{R}$ for all T. In contrast with the setting in Section 4.3, we define the new measure as $\mu(dt) = \frac{h(t,r(t))^2}{\sigma(t,r(t))^2}dt$ on $[0,T]$. The continuity of $h(\cdot)$ and $\{r(t), 0 \le t \le T\}$ implies that $\int_0^T \mu(dt) < +\infty$ for any finite T. We take $\vec{p}_{T,K}(t) = \vec{q}_{T,K}(t)\frac{\sigma(t,r(t))}{h(t,r(t))}$. Then $A_{T,K} = \frac{T}{2}I_{(K+1)\times(K+1)}$, and analytical results can be derived in a manner similar to that described in Section 4.3.

In the following, we derive the maximum likelihood estimator of $M_{T,K}(\theta)(t)$. Since $\{\theta(t)\}$ is bounded on $[0,T]$ and hence lies within the space $L^2([0,T], \mu(dt))$, we can write

$$\theta(t) = M_{T,K}(\theta)(t) + M^{\perp}_{T,K}(\theta)(t) = \vec{p}_{T,K}(t)'\Phi_{T,K}(\theta) + M^{\perp}_{T,K}(\theta)(t), \quad 0 \le t \le T.$$

Since $\{p_{i,T}\}_{i\ge 0}$ are linearly independent functions on $[0,T]$, we have a unique representation of $\Phi_{T,K}(\theta) = (\theta_{0,T}, \theta_{1,T}, \cdots, \theta_{K,T})'$. Therefore, we have the following

representation:

$$\theta(t) = \sum_{j=0}^{K} \theta_{j,T} p_{j,T}(t) + M_{T,K}^{\perp}(\theta)(t), \quad 0 \le t \le T.$$

Let P_r^T be the probability measure generated by the process $\{r(t) : 0 \le t \le T\}$ on the space $(C[0,T], \mathcal{B}_T)$. Let P_η^T be the probability measure induced by the strong solution to the equation: $d\eta(t) = \sigma(t, \eta(t))dW(t), \quad 0 \le t \le T$. Under the additional conditions

$$P(\omega \in \Omega : |h(t, r(t))| < \infty) = 1, \quad \forall\ 0 \le t \le T \tag{4.25}$$

$$P(\omega \in \Omega : \int_0^T h^2(t, r(t))dt < \infty) = 1, \tag{4.26}$$

we have $P_r^T \ll P_\eta^T$ (Lipster & Shiryaev (1974)) and

$$\begin{aligned}
&\frac{dP_r^T}{dP_\eta^T} \\
= \ & \exp\{\int_0^T \frac{(\sum_{j=0}^K \theta_{j,T} p_{j,T}(t) + M_{T,K}^{\perp}(\theta)(t))h(t, r(t))}{\sigma^2(t, r(t))} dr(t) - \\
& \frac{1}{2}\int_0^T \frac{(\sum_{j=0}^K \theta_{j,T} p_{j,T}(t) + M_{T,K}^{\perp}(\theta)(t))^2 h^2(t, r(t))}{\sigma^2(t, r(t))} dt\} \\
= \ & \exp\{\int_0^T \frac{(\sum_{j=0}^K \theta_{j,T} p_{j,T}(t))h(t, r(t))}{\sigma^2(t, r(t))} dr(t) + \int_0^T \frac{(M_{T,K}^{\perp}(\theta)(t))h(t, r(t))}{\sigma^2(t, r(t))} dr(t) - \\
& \frac{1}{2}\int_0^T \frac{(\sum_{j=0}^K \theta_{j,T} p_{j,T}(t))^2 h^2(t, r(t))}{\sigma^2(t, r(t))} dt - \frac{1}{2}\int_0^T \frac{(M_{T,K}^{\perp}(\theta)(t))^2 h^2(t, r(t))}{\sigma^2(t, r(t))} dt\},
\end{aligned}$$

where the last equality holds because $M_{T,K}^{\perp}(\theta)$ and $p_{j,T}, 0 \le j \le K$ are orthogonal within $L^2([0,T], \mu(dt))$. By treating $M_{T,K}^{\perp}(\theta)(t)$ as a nuisance parameter, the log-likelihood function of $\Phi_{T,K}(\theta) = (\theta_{0,T}, \cdots, \theta_{K,T})'$ can be derived as

$$\begin{aligned}
& l(\Phi_{T,K}(\theta)) \\
= \ & \int_0^T \frac{h(t, r(t))\vec{p}'_{T,K}\Phi_{T,K}(\theta)}{\sigma^2(t, r(t))} dr(t) + \int_0^T \frac{h(t, r(t))M_{T,K}^{\perp}(\theta)(t)}{\sigma^2(t, r(t))} dr(t) - \\
& \frac{1}{2}\int_0^T \frac{(\vec{p}'_{T,K}\Phi_{T,K}(\theta))^2 h^2(t, r(t))}{\sigma^2(t, r(t))} dt - \frac{1}{2}\int_0^T \frac{(M_{T,K}^{\perp}(\theta)(t))^2 h^2(t, r(t))}{\sigma^2(t, r(t))} dt.
\end{aligned}$$

It follows that the score function for $\Phi_{T,K}(\theta)$ can be calculated:

$$S(\Phi_{T,K}(\theta)) = \int_0^T \frac{h(t,r(t))\vec{p}_{T,K}(t)}{\sigma^2(t,r(t))}dr(t) - \int_0^T \frac{\vec{p}_{T,K}(t)\vec{p}'_{T,K}(t)\Phi_{T,K}(\theta)h^2(t,r(t))}{\sigma^2(t,r(t))}dt.$$

By solving $S(\hat{\vec{\theta}}_{T,K}) = 0$, we obtain

$$\hat{\Phi}_{T,K}(\theta) = A_{T,K}^{-1}\int_0^T \frac{h(t,r(t))\vec{p}_{T,K}(t)}{\sigma^2(t,r(t))}dr(t) = A_{T,K}^{-1}\int_0^T \frac{\vec{q}_{T,K}(t)}{\sigma(t,r(t))}dr(t),$$

where

$$A_{T,K} \triangleq \int_0^T \frac{\vec{p}_{T,K}(t)\vec{p}'_{T,K}(t)h^2(t,r(t))}{\sigma^2(t,r(t))}dt = \int_0^T \vec{q}_{T,K}(t)\vec{q}'_{T,K}(t)dt = \frac{T}{2}I_{(K+1)\times(K+1)}.$$

We can apply now the same techniques as in the previous sections to derive some properties of the maximum likelihood estimator $\hat{\Phi}_{T,K}(\theta)$ and $\hat{M}_{T,K}(\theta)(t) \triangleq \vec{p}'_{T,K}(t)\hat{\Phi}_{T,K}(\theta)$. Below we list the main results with some comments:

1.
$$\hat{\Phi}_{T,K}(\theta) \sim MVN(\Phi_{T,K}(\theta), (A_{T,K})^{-1}).$$

2. For each fixed t,
$$\frac{h(t,r(t))(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(\theta)(t))}{\sigma(t,r(t))} \sim N(0, \vec{q}_{T,K}(t)'(A_{T,K})^{-1}\vec{q}_{T,K}(t)).$$

3. The 95% confidence interval for the projection $M_{T,K}(\theta)(t)$ is
$$\left(\hat{M}_{T,K}(\theta)(t) - z_{0.975}\frac{\sigma(t,r(t))}{h(t,r(t))}\sqrt{\vec{q}_{T,K}(t)'(A_{T,K})^{-1}\vec{q}_{T,K}(t)},\right.$$
$$\left.\hat{M}_{T,K}(\theta)(t) - z_{0.025}\frac{\sigma(t,r(t))}{h(t,r(t))}\sqrt{\vec{q}_{T,K}(t)'(A_{T,K})^{-1}\vec{q}_{T,K}(t)}\right),$$
or equivalently,
$$\left(\hat{M}_{T,K}(\theta)(t) - z_{0.975}\sqrt{\vec{p}_{T,K}(t)'(A_{T,K})^{-1}\vec{p}_{T,K}(t)},\right.$$
$$\left.\hat{M}_{T,K}(\theta)(t) - z_{0.025}\sqrt{\vec{p}_{T,K}(t)'(A_{T,K})^{-1}\vec{p}_{T,K}(t)}\right),$$

where $z_{0.975}$ and $z_{0.025}$ are the quantiles of a standard normal random variable. From the form of the interval, we can infer that when $\frac{h(t,r(t))}{\sigma(t,r(t))}$ is close to zero, the estimation for $M_{T,K}(\theta)(t)$ is very unreliable due to the enormous variance of $\hat{M}_{T,K}(\theta)(t)$. This finding generalizes similar comments made by Nguyen and Pham (1982), who studied the simpler model (4.23) under the assumption that we observe many independent realizations of the process.

4. Replacing $\sigma(t, r(t))$ in Section 4.3.4 by $\frac{\sigma(t,r(t))}{h(t,r(t))}$, we can define a similar hypothesis-testing procedure as before.

5. $\hat{M}_{T,K_T}(\theta)(t)$ is a weakly consistent estimator of $M_{T,K_T}(\theta)(t)$ under the following condition:

   a) $\lim_{T\to\infty} \frac{K_T^2}{T} = 0$ if $\{\vec{R}(t) : -1 \leq t \leq 1\}$ are the normalized Legendre polynomials.

   b) $\lim_{T\to\infty} \frac{K_T}{T} = 0$ if $\{\vec{R}(t) : -1 \leq t \leq 1\}$ are the normalized trigonometric polynomials.

4.5 Simulation Studies

In this section, we illustrate our methodology using one simulated trajectory of 10 years of daily data from the model

$$dr(t) = \theta(t)dt + \sigma dW(t), \quad 0 \leq t \leq T,$$

where σ is a known constant. The objective is to estimate $\{M_{T,K}(\theta)(t)\}$. We choose the basis system $\vec{p}_T(t) = \sigma\vec{q}_T(t)$, where $\vec{q}_T(t)$ are Legendre polynomials on $[0, T]$, as defined in Appendix 4.7.2. The chosen size of the hypothesis test is 5%. We give examples for both a smooth drift function, which is a polynomial function, and non-smooth drift function, which is a piecewise polynomial function. The following estimation method is employed in our simulation studies:

- The maximum likelihood estimator $\hat{M}_{T,K}(\theta)(t)$ is based on (4.8):

$$\hat{M}_{T,K}(\theta)(t) = \vec{p}'_{T,K}(t)A^{-1}_{T,K}\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma^2}dr(t) = \frac{2}{T}\vec{q}'_{T,K}(t)\int_0^T \vec{q}_{T,K}(t)dr(t).$$

- The confidence interval for $M_{T,K}(\theta)(t)$ is given by (4.17) in Section 4.3.5:

$$\left(\hat{M}_{T,K}(\theta)(t) - z_{0.975}\sigma(t,r(t))\sqrt{\frac{2}{T}\vec{q}_{T,K}(t)'\vec{q}_{T,K}(t)},\right.$$
$$\left.\hat{M}_{T,K}(\theta)(t) - z_{0.025}\sigma(t,r(t))\sqrt{\frac{2}{T}\vec{q}_{T,K}(t)'\vec{q}_{T,K}(t)}\right).$$

- The hypothesis-testing procedure is from Section 4.3.4 and the χ^2 test statistics is from (4.16).

4.5.1 The Smooth Drift Case

In the following, we set $\theta(t)$ to be a smooth function equal to 0.0672 $L(2,10)$, where $L(2,10)$ denotes a Legendre polynomial of degree 2 on interval $[0,10]$. We illustrate firstly our estimation method with different σ and K_T, then the hypothesis testing procedure to determine the dimension of parameter space, and lastly the selection of a subset of basis functions to minimize the IMSE of $\hat{M}_{T,K}(\theta)$.

Now we consider two examples: one with a smaller diffusion coefficient and the other with a larger value. First, we set $\sigma = 0.0224$. In this case, the drift to diffusion ratio is approximately 3 to 1. Figure 4.3 plots one simulated realization of the process.

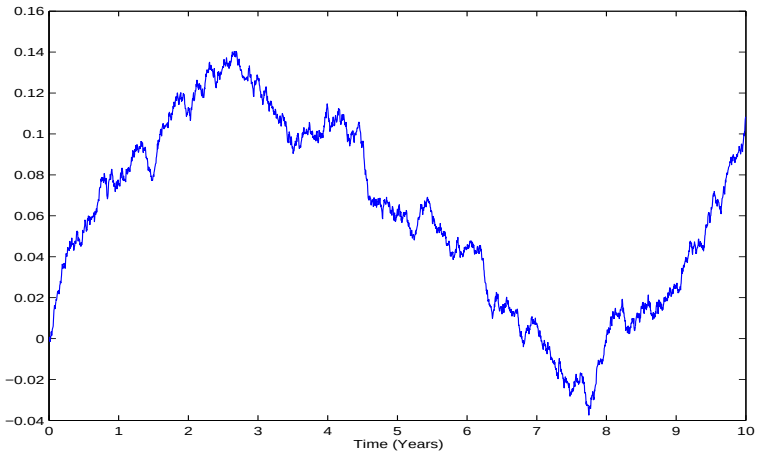


Figure 4.3: Simulated BM: $\theta(t) = 0.0672\ L(2, 10)$, $\sigma = 0.0224$, $K = 2$.

Figure 4.4 shows the estimated drift function and its confidence interval for $K = 2$. Figure 4.5 (a) and (b) show the estimation results when we use $K = 0$ and $K = 5$. By comparing these figures, one can see that the confidence interval for the projection of the drift function $M_{T,K}(\theta)$ is larger when more basis functions are incorporated.

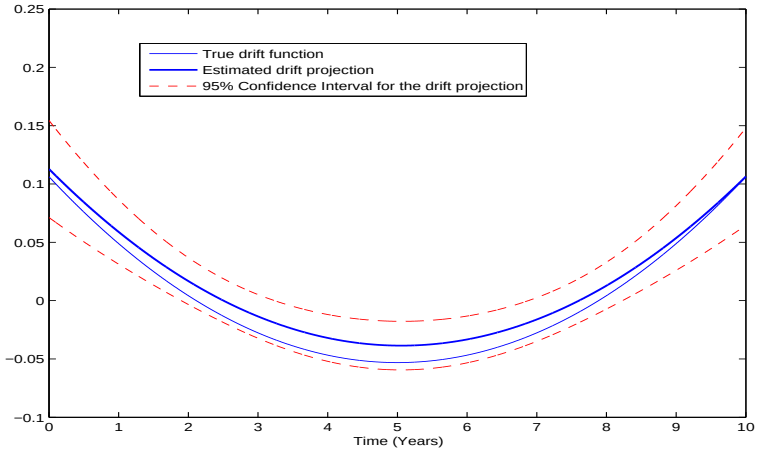
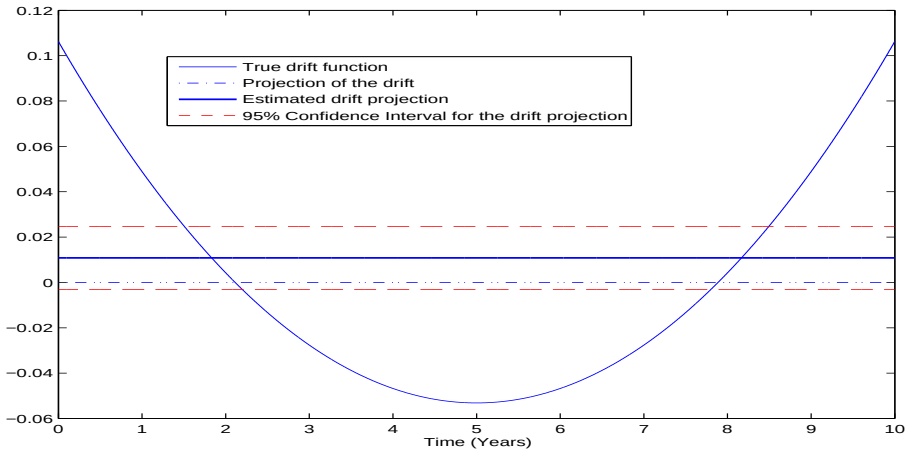
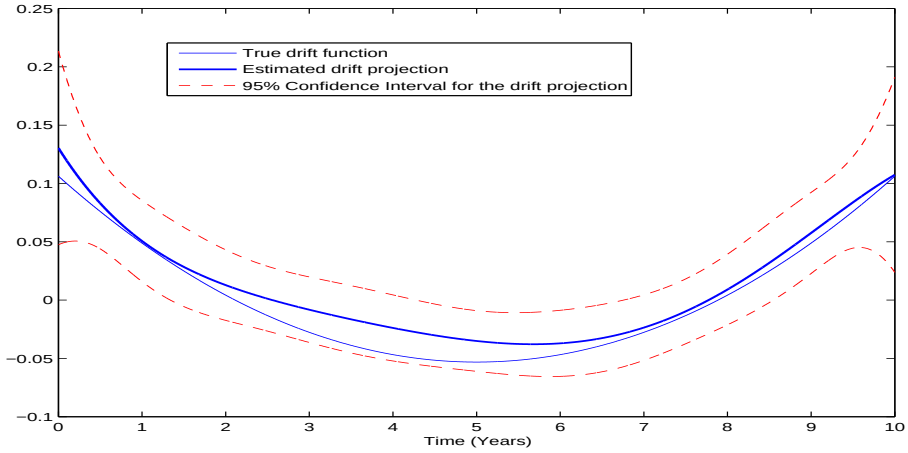


Figure 4.4: Drift estimation for BM: $\theta(t) = 0.0672\ L(2, 10)$, $\sigma = 0.0224$, $K = 2$



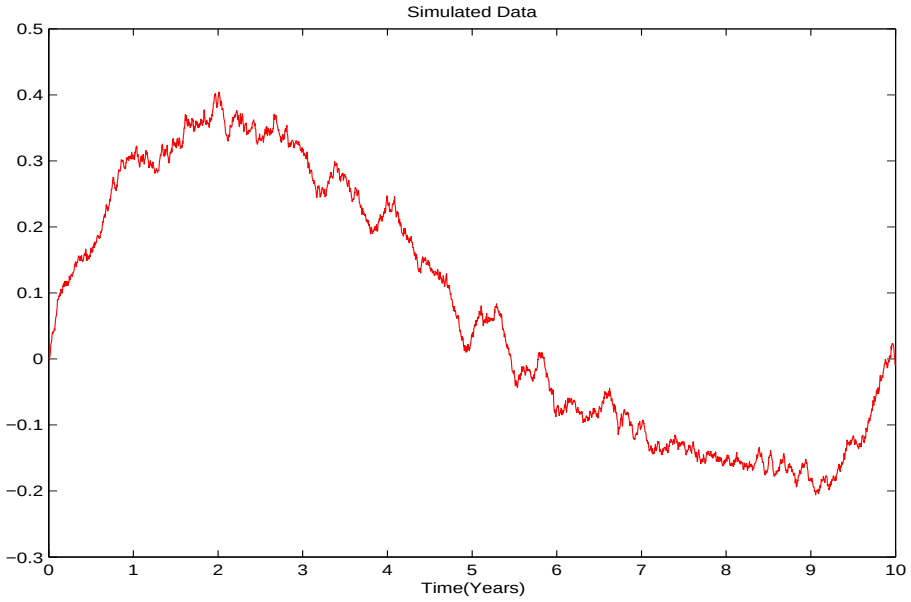
(a)



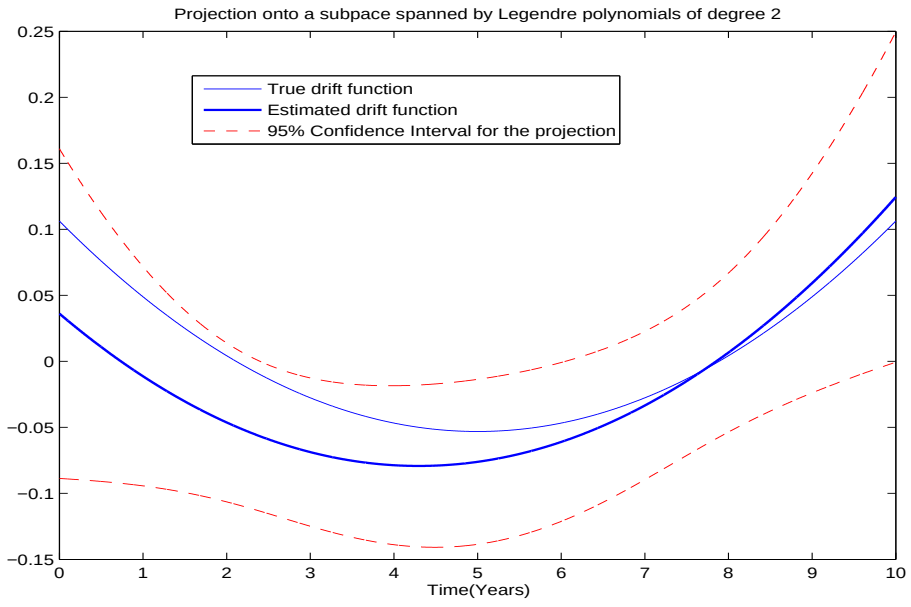
(b)

Figure 4.5: Drift estimation for BM: $\theta(t) = 0.0672\ L(2, 10)$, $\sigma = 0.0224$, $K = 0$ or 5.

Next, we set $\sigma = 0.0672$; i.e., the drift-to-diffusion ratio is approximately 1 to 1. Figure 4.6 (a) plots a simulated Brownian motion, and Figure 4.6 (b) shows the estimated drift function and its confidence interval for $K = 2$. In contrast with the previous case when $\sigma = 0.024$, now the confidence interval for the projection $M_{T,K}(\theta)$ is significantly larger because the drift-to-diffusion ratio has decreased from 3 to 1.



(a)



(b)

Figure 4.6: Drift estimation for BM: $\theta(t) = 0.0672\ L(2, 10)$, $\sigma = 0.0672$, $K = 2$

Now, we demonstrate that the proposed hypothesis-testing method can be used to detect whether we are fitting the data with too few or too many basis functions. $K_{max} = 20$ is chosen for this purpose, and the chi-square statistic derived from Section 4.3.4 is employed. Table 4.1 summarizes the hypothesis-testing results on the true degree of the time-dependent drift function. Based on the result, we cannot reject the null hypothesis that the degree of $\theta(t)$ is no larger than 2. Therefore, the dimension of $\theta(t)$ is correctly detected by the proposed hypothesis testing method.

Table 4.1: Hypothesis-testing results for a smooth drift function

H_0: the degree of $\theta(t)$ is no larger than	0	1	2	3	4	5	6	7	8	9	10
p-value (%)	0.0	0.0	86.0	90.8	88.3	84.2	78.8	79.5	72.8	74.5	73.4

Now we apply our proposed method to select the subset of Legendre polynomials that minimizes the IMSE, which we describe in Section 4.3.6. In this simulation study, we choose $K_{max} = 20$. According to the three steps we have proposed, we first estimate all coefficients corresponding to Legendre polynomials of degree up to 20. Then we sort the estimated coefficients in a descending order and choose the polynomials with coefficient estimates larger than $\frac{1}{\sqrt{5}}$, which corresponds to taking $\sqrt{\frac{2}{T}}$ in Step 1 in Section 4.3.6. Figure 4.7 shows the estimation result. The selected Legendre polynomials are of degree $2, 11, 3, 16$ and 9, selected in a descending magnitude of coefficient estimates. Since $\theta(t) = 0.0672\ L(2, 10)$, the simulation result has successfully identified that the key Legendre polynomial to explain $\theta(t)$ is of degree 2.

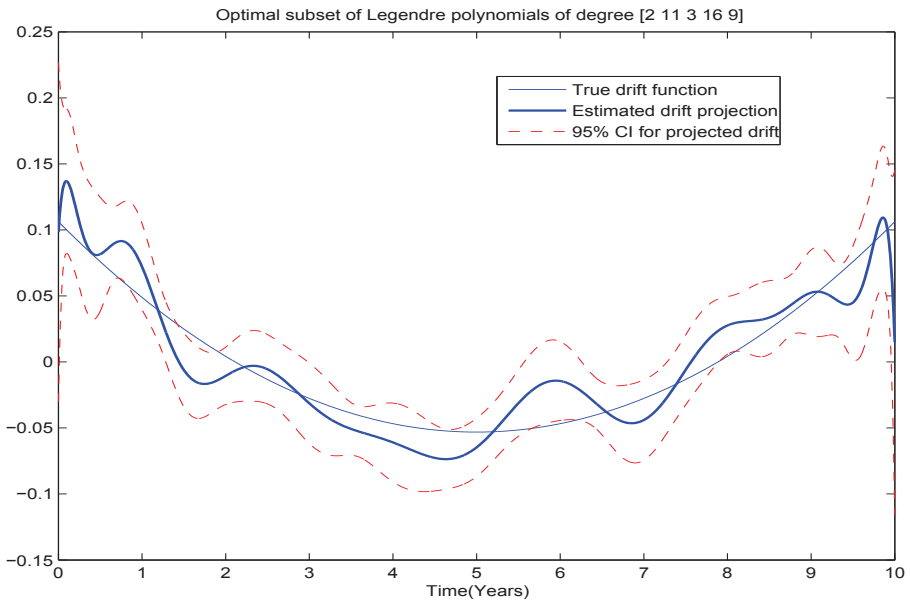


Figure 4.7: Drift estimation for BM: $\theta(t) = 0.0672\ L(2, 10)$, $\sigma = 0.0224$. The selected subset includes Legendre polynomials of degree $2, 11, 3, 16$ and 9, in a descending magnitude of coefficient estimates.

4.5.2 The Non-Smooth Drift Case

We now simulate a Brownian motion with a piecewise polynomial drift function and $\sigma = 0.0224$. We define each piece of the drift as a quadratic function, and the drift-to-diffusion ratio is approximately 3 to 1. The piecewise polynomial function is of the form

$$\theta(t) = \begin{cases} \frac{3\sigma^2}{10} & 0 \leq t \leq 4 \\ -\frac{3\sigma(t-4)^2}{10} & 4 < t \leq 7 \\ \frac{3\sigma(t-7)^2}{10} & 7 < t \leq 10 \\ -\frac{3\sigma(t-10)^2}{10} & 10 < t \leq 20 \end{cases}$$

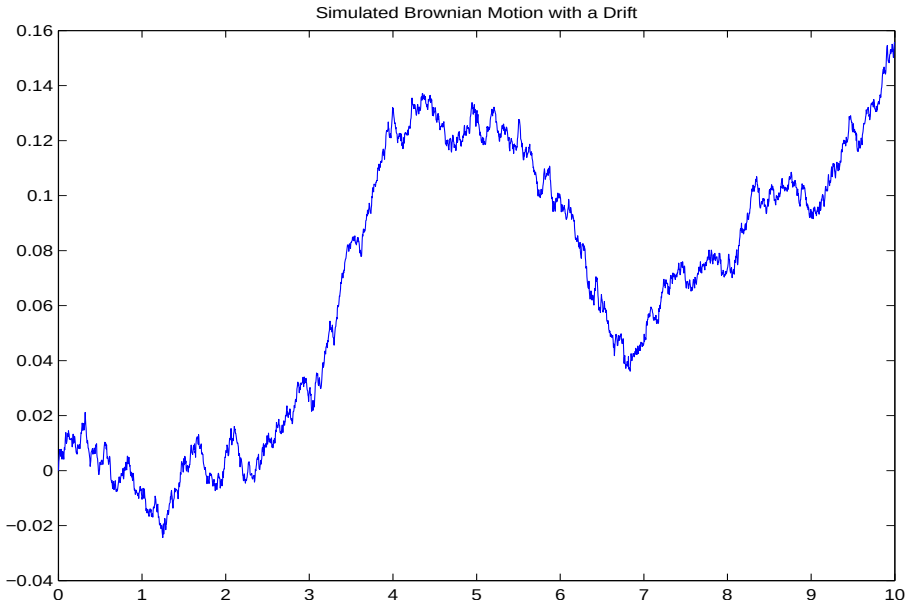
As before, we use Legendre polynomials as our basis functions. We first conduct our proposed hypothesis-testing procedure to determine the dimension of the projection space. The subspace $V_{T,K}$ used for projections is spanned by Legendre polynomials of degree from 0 to $K_{max} = 20$. Table 4.2 shows p-values for our hypothesis testing when the degree of Legendre polynomial is at most 10. When the degree is greater than 10, the p-values are greater than 5%, and we omit them here. If the p-value is smaller than 5%, the null hypothesis is rejected. From these results, $K_T = 0, 1, \cdots, 4$ are rejected by our method, and $K_T = 5$ is employed to estimate the projection of the drift function.

Table 4.2: Hypothesis test for BM with non-smooth drift function

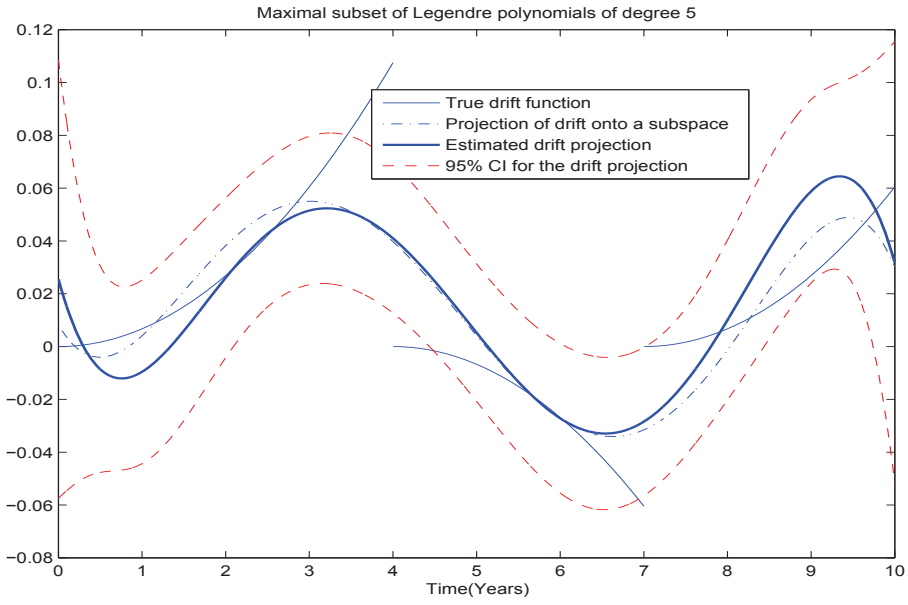
H_0: the degree of $\theta(t)$ is no larger than	0	1	2	3	4	5	6	7	8	9	10
p-value (%)	0.2	0.2	0.1	1.8	1.2	5.8	24.7	30.9	24.3	50.0	45.6

Figure 4.8 shows the simulated data and estimation results. The projection of the true drift function onto $V_{T,K=5}$ is within our 95% confidence interval, supporting that our proposed method of estimation and hypothesis testing is useful in estimating a projection of the non-smooth drift function onto a finite dimensional space.

We also demonstrate an application of the method for selecting the optimal subset of Legendre polynomials, which minimizes the IMSE. Figure 4.9 shows the estimation result. Using the three steps described in Section 4.3.6, we have selected Legendre polynomials of the following degrees $3, 10, 5, 20, 13, 6, 18, 7, 0, 1, 14, 8$ and 11, which are listed in a descending magnitude of coefficients in the expansion(4.20). The estimation accuracy in Figures 4.9 and 4.8 seems similar.



(a)



(b)

Figure 4.8: BM with a piecewise continuous drift, $K = 5$.

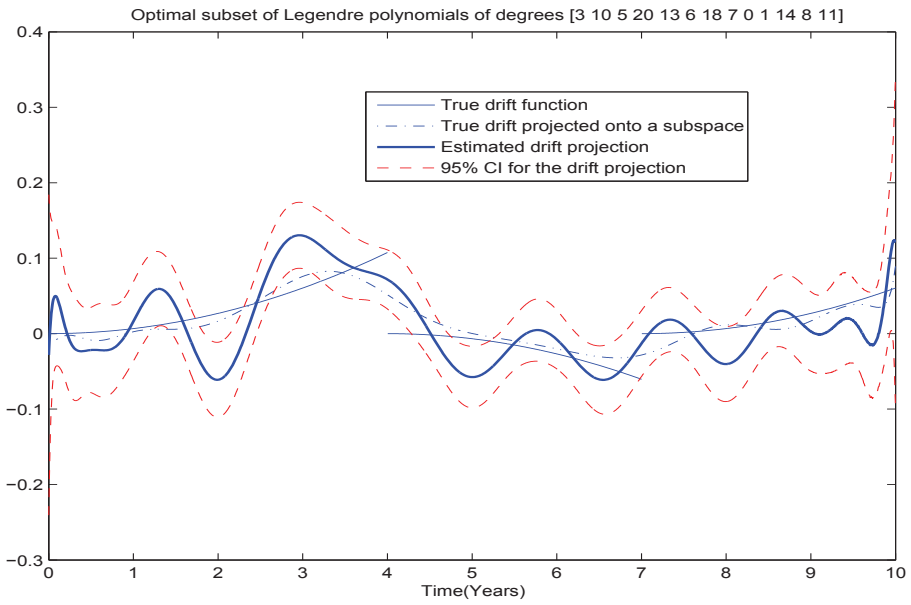


Figure 4.9: Drift estimation for BM with non-smooth drift: $\sigma = 0.0224$. The selected subset includes Legendre polynomials of degrees $3, 10, 5, 20, 13, 6, 18, 7, 0, 1, 14, 8$ and 11, in a descending magnitude of coefficient estimates.

4.6 Concluding Remarks

In this chapter, we have studied a stochastic differential equation with time-dependent drift functions. Based on one single continuous realization, our objective is to estimate the projection of the time-dependent drift onto $V_{T,K}$, a subspace in $L^2([0,T],\mu(dt))$. We have derived the closed-form expression of the maximum likelihood estimator of the projected drift. Moreover, we prove the asymptotic consistency of the estimator in a sieve type setting.

The key requirement in proving the consistency result in Section 4.3.3 is that the second term on the right-hand side of the equation (4.10) converges to zero

in probability. The main challenge is that $A_{T,K}^{-1}$ can be very involved, making the analysis difficult. We solve this problem by choosing the basis functions $\vec{p}_{T,K}(t)$ in a particular way such that $A_{T,K}^{-1} = \frac{T}{2} I_{(K+1)\times(K+1)}$. The proof then becomes much easier in Section 4.3.3.

The proposed method for constructing a sequence of maximum likelihood estimators over restricted parameter space can be extended to more general time-inhomogeneous SDEs. However, we do not expect an exact distribution of the maximum likelihood estimator. In Chapter 5, we study a time-dependent component in a mean-reverting SDE. The mean-reverting feature introduces the aliasing problem of estimating the mean reverting speed parameter and the time-dependent level function. The unknown distribution of the maximum likelihood estimator is the main challenge in extending the methods proposed in this chapter.

4.7 Appendix

4.7.1 Chebyshev Polynomials

The Chebyshev discrete time polynomials take the following form:

$$p_{0,n}(t) = 1, \quad p_{j,n(t)} = \sqrt{2} cos[\frac{j\pi(t-0.5)}{n}], \quad j = 1, \cdots, n-1, t = 1, \cdots, n.$$

It can be shown (Hamming 1973, Bierens 1997) that for $k, m = 0, 1, \cdots, n-1$, we have $(1/n)\sum_{t=1}^{n} p_{k,n}(t) p_{m,n}(t) = I(k = m)$, where $I(\cdot)$ is the indicator function.

4.7.2 Legendre Polynomials and Trigonometric Polynomials

Legendre Polynomials

Let $\{l_n(t):\ n=0,1,2,\cdots\}$ be the Legendre polynomials on $[-1,1]$:

$$l_n(t)=(-1)^n\sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right)\left(\begin{array}{c} n+k \\ k \end{array}\right)(-\frac{t+1}{2})^k.$$

Some well-known facts about these polynomials include that

- They form an orthogonal set with respect to Lebesgue measure.
- $|l_n(t)|\leq 1,\ -1\leq t\leq 1$.
- $\int_{-1}^{1} l_n^2(t)dt=\frac{2}{2n+1}$.
- $l_n(-1)=(-1)^n$ and $l_n(1)=1,\ \forall\ n=0,1,\cdots$
- $\tilde{l}_n(t)\triangleq l_n(2t-1),\ 0\leq t\leq 1$, are called the shifted Legendre polynomials on $[0,1]$.

We call $R_n(t)\triangleq\sqrt{\frac{2n+1}{2}}l_n(t)$ the normalized Legendre polynomials on $[-1,1]$, due to the fact that $\int_{-1}^{1}R_n^2(t)dt=1,\ \forall\ n=0,1,\cdots$. Moreover, we have $|R_n(t)|\leq\sqrt{\frac{2n+1}{2}},\ -1\leq t\leq 1$. Figure 4.10 depicts shifted and scaled Legendre polynomials $\{R_n(2t-1),0\leq t\leq 1\}$. They are the same as $\vec{q}_{T=1,K},0\leq K\leq 6$ introduced in Section 4.3.1. Since $q_{i,T}(t)=R_i(\frac{2t}{T}-1)$, it follows from elementary calculus that $\int_0^T q_{i,T}^2(t)dt=\frac{T}{2},\ \forall\ i\geq 0$.

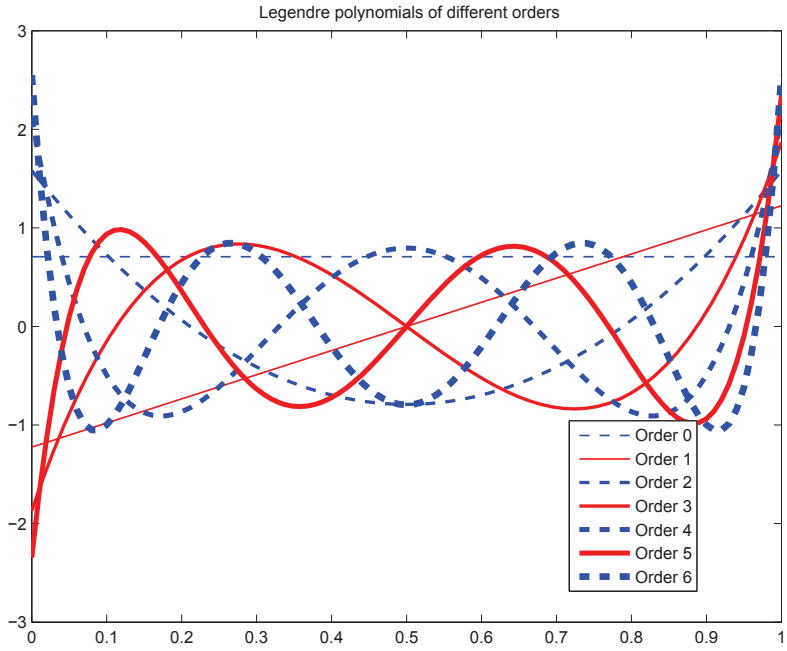


Figure 4.10: Orthogonal Legendre polynomials

Trigonometric polynomials

Trigonometric polynomials on $[-1, 1]$ are of the form $\{1, cos(\pi t), sin(\pi t), cos(2\pi t), sin(2\pi t), \cdots\}$. They form an orthogonal set:

$$\begin{aligned}
\int_{-1}^{1} cos(m\pi t)sin(n\pi t)dt &= \frac{1}{2}sin[(m+n)\pi t]\ |_{-1}^{1} - \frac{1}{2}sin[(m-n)\pi t]|\ |_{-1}^{1} = 0 \\
\int_{-1}^{1} cos(m\pi t)cos(n\pi t)dt &= \frac{1}{2}cos[(m+n)\pi t]|_{-1}^{1} + \frac{1}{2}cos[(m-n)\pi t]|_{-1}^{1} = \delta_{mn} \\
\int_{-1}^{1} sin(m\pi t)sin(n\pi t)dt &= -\frac{1}{2}cos[(m+n)\pi t]|_{-1}^{1} + \frac{1}{2}cos[(m-n)\pi t]|_{-1}^{1} = \delta_{mn},
\end{aligned}$$

where $m, n = 0, 1, 2, \cdots$ and δ_{mn} is the Kronecker delta. The above trigonometric polynomials have a unit norm in $L^2([-1, 1], dt)$ and hence are normalized trigonometric polynomials. In this case, we have $\vec{R}(t) = \{1, cos(\pi t), sin(\pi t), cos(2\pi t), sin(2\pi t), \cdots\}$. With $q_{i,T}(t)$ as introduced in Section

4.3.1, it follows from elementary calculus that $\int_0^T q_{i,T}^2(t)dt = \frac{T}{2}$, $\forall\ i \geq 0$.

4.7.3 Derivation for the Projection Operator $M_{T,K}$

For any $\{u(t) : 0 \leq t \leq T\} \in L^2([0,T], \mu(dt))$, we can express the projection of $\{u(t) : 0 \leq t \leq T\}$ onto $V_{T,K}$ as $M_{T,K}(u)(t) = \sum_{j=0}^{K} u_{j,T}\ p_{j,T}(t)$, where $u_{j,T} \in \mathbb{R}$. By the definition of orthogonal projection, $< u - M_{T,K}(u), p_{j,T} >_\mu = 0, \forall\ j \in \{0, \cdots, K\}$. Therefore,

$$< u - \vec{p'}_{T,K}\vec{u}_{T,K}, p_{i,T} >_\mu = 0, \forall\ i \in \{0, \cdots, K\}, \tag{4.27}$$

where $\vec{u}_{T,K} \triangleq (u_{0,T}, u_{1,T}, \cdots, u_{K,T})'$. It follows from (4.27) that

$$\int_0^T u(t)p_{i,T}(t)\mu(dt) = \int_0^T p_{i,T}(t)\vec{p}_{T,K}(t)'\vec{u}_{T,K}\mu(dt),$$

which implies

$$\int_0^T u(t)\vec{p}_{T,K}(t)\mu(dt) = \int_0^T \vec{p}_{T,K}(t)\vec{p}_{T,K}(t)'\vec{u}_{T,K}\mu(dt). \tag{4.28}$$

Define $A_{T,K} \triangleq \int_0^T \vec{p}_{T,K}(t)\vec{p}_{T,K}(t)'\mu(dt)$; then (4.28) becomes

$$\vec{u}_{T,K} = A_{T,K}^{-1} \int_0^T u(t)\vec{p}_{T,K}(t)\mu(dt),$$

assuming that $A_{T,K}^{-1}$ exists. Therefore,

$$M_{T,K}(u)(t) = \vec{p}_{T,K}(t)' A_{T,K}^{-1} \int_0^T u(t)\vec{p}_{T,K}(t)\mu(dt).$$

4.7.4 Positive Definiteness of $HA_{T,K_{max}}^{-1}H'$

First we show the positive definiteness of $A_{T,K}$ for any $K \geq 0$. For any row vector $c = (c_0, c_1, \cdots, c_K) \in \mathbb{R}^{K+1}$, we have

$$\begin{aligned} cA_{T,K}c' &= c\left(\int_0^T \vec{p}_{T,K}(t)\vec{p}_{T,K}(t)'\mu(dt)\right)c' \\ &= \int_0^T (c\vec{p}_{T,K}(t))\ (c\vec{p}_{T,K}(t))'\,\mu(dt) \\ &= \int_0^T \left(\sum_{i=0}^K c_i p_{i,T}(t)\right)^2 \mu(dt). \end{aligned}$$

Since $\mu(dt) = \frac{1}{\sigma^2(t,r(t))}dt$, the positivity and boundedness of $\sigma(t, r(t))$ imply that $\int_0^T (\sum_{i0}^K c_i p_{i,T}(t))^2 \mu(dt) \geq 0$ for any $c \in \mathbb{R}^{K+1}$. Moreover, $\int_0^T (\sum_{i=0}^K c_i p_{i,T}(t))^2 \mu(dt) = 0$ holds if and only if $\sum_{i=0}^K c_i p_{i,T}(t) = 0$ almost surely with respect to measure μ. Then the linear independence of $p_{i,T}, i \geq 0$ implies that $c_i = 0, \forall i \in \{0, \cdots, K\}$. Therefore, we have proven the positive definiteness of $A_{T,K}$. One byproduct of the positive definiteness of $A_{T,K}$ is that it is an invertible matrix.

In the following, we prove the positive definiteness of $HA_{T,K_{max}}^{-1}H'$. For any row vector $l = (l_1, l_2, \cdots, l_{K_{max}-K_0+1}) \in \mathbb{R}^{K_{max}-K_0+1}$, we have

$$lH = (0, \cdots, 0, l_1, l_2, \cdots, l_{K_{max}-K_0+1}),$$

where the first K_0 elements of lH are 0, and lH is a 1 by $K_{max}+1$ row vector. Since the positive definiteness of $A_{T,K}$ implies that $A_{T,K}^{-1}$ is also positive definite, we have $lHA_{T,K}^{-1}H'l' \geq 0$ for any $l \in \mathbb{R}^{K_{max}-K_0+1}$. Moreover, if $lHA_{T,K}^{-1}H'l' = 0$, then $lH = 0$. This implies that $l_i = 0, \forall\ i \in \{1, \cdots, K_{max} - K_0 + 1\}$.

4.7.5 Technical Proofs

Proof of Theorem 4.2

(i) From (4.12), we have $\frac{\hat{M}_{T,K_T}(\theta)(t)-M_{T,K_T}(\theta)(t)}{\sigma(t,r(t))} \sim N(0, \frac{2}{T}\sum_{i=0}^{K_T} q_{i,T}^2(t))$ for each t and T. If $K_T = K$; that is, K_T does not increase with T, then

$$\begin{aligned}\sqrt{T}\left(\frac{\hat{M}_{T,K}(\theta)(t)-M_{T,K}(\theta)(t)}{\sigma(t,r(t))}\right) &\sim N(0, 2\sum_{i=0}^{K} q_{i,T}^2(t)) \\ &= N(0, 2\sum_{i=0}^{K} R_i^2(\frac{2t}{T}-1)).\end{aligned}$$

By the continuity of $R_i(\cdot)$ we have $\lim_{T\to\infty} R_i(\frac{2t}{T}-1) = R_i(-1)$. Then

$$\sqrt{T}\left(\frac{\hat{M}_{T,K}(\theta)(t)-M_{T,K}(\theta)(t)}{\sigma(t,r(t))}\right) \xrightarrow{D} N(0, 2\sum_{i=0}^{K} R_i^2(-1)), \ \ as\ T\to\infty, \quad (4.29)$$

where $\xrightarrow{D}$ means convergence in distribution. Since

$$\hat{M}_{T,K}(\theta)(t)-M_{T,K}(\theta)(t) = \frac{\sigma(t,r(t))}{\sqrt{T}} \times \sqrt{T}\left(\frac{\hat{M}_{T,K}(\theta)(t)-M_{T,K}(\theta)(t)}{\sigma(t,r(t))}\right),$$

it follows from the boundedness of $\sigma(t,r(t))$ and (4.29) that $\hat{M}_{T,K}(\theta)(t) - M_{T,K}(\theta)(t)$ converges to zero in probability as $T\to\infty$. Therefore, for each $0\le t\le T$, $\hat{M}_{T,K}(\theta)(t)$ is a weakly consistent estimator of $M_{T,K}(\theta)(t)$.

(ii) If K_T changes with T, we investigate two situations: when $\{\vec{R}(t) : -1 \le t \le 1\}$ are either normalized Legendre polynomials or normalized trigonometric polynomials:

(1) If $\{\vec{R}(t) : -1 \le t \le 1\}$ are the normalized Legendre polynomials, then $q_{i,T}^2(t) \le \frac{2i+1}{2}$ from Appendix 4.7.2. Therefore,

$$\frac{2}{T}\sum_{i=0}^{K_T} q_{i,T}^2(t) \le \frac{2}{T}\sum_{i=0}^{K_T}\frac{2i+1}{2} = \frac{1}{T}(K_T+1)^2. \quad (4.30)$$

Under the assumption $\lim_{T\to\infty}\frac{K_T^2}{T} = 0$, it follows from (4.12) and Chebyshev's inequality that for each fixed t, $\frac{\hat{M}_{T,K_T}(\theta)(t)-M_{T,K_T}(\theta)(t)}{\sigma(t,r(t))}$ converges to zero in probability as $T\to\infty$. Since for model (4.1) we have

assumed $0 < \epsilon < \sigma(t, r(t)) < M$, the consistency of $\hat{M}_{T,K_T}(\theta)(t)$ follows from Slutsky's theorem.

(2) If $\{\vec{R}(t) : -1 \leq t \leq 1\}$ are the normalized trigonometric polynomials, then $R_i^2(t) \leq 1$ (Appendix 4.7.2). Then

$$\frac{2}{T}\sum_{i=0}^{K_T} q_{i,T}^2(t) \leq \frac{1}{T}(K_T + 1). \tag{4.31}$$

Under the assumption $\lim_{T\to\infty} \frac{K_T}{T} = 0$, it follows from (4.12) and Chebyshev's inequality that for each fixed t, $\dfrac{\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)}{\sigma(t, r(t))}$ converges to zero in probability and $T \to \infty$. A similar argument to the one above then leads to the conclusion that $\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)$ converges to zero as $T \to \infty$. □

Proof of Theorem 4.4

Let F_N denote a subset of $\mathbb{N} = \{0, 1, 2, \cdots\}$ with cardinality N. Our objective is to prove that $IMSE(\hat{\theta}_{T,K_{opt}}(\cdot)) \leq IMSE(\hat{\theta}_{F_N}(\cdot))$ for any $N \geq 0$.

For any $N > K_{max} + 1$, we first show that there exists a positive integer N^* such that $N^* \leq K_{max} + 1$ and $IMSE(\hat{\theta}_{F_{N^*}}(\cdot)) \leq IMSE(\hat{\theta}_{F_N}(\cdot))$.

Assume that $F_N = \{j_1, j_2, \cdots, j_N\}$ and delete all the j_ks that are greater than K_{max}, where the indices refer to the given complete orthogonal basis system $\{p_{0,T}, p_{1,T}, \cdots\}$. We denote this new set by F_{N^*}. The cardinality of F_{N^*} will be at most K_{max}. Similar to (4.21), we have

$$\begin{aligned} IMSE(\hat{\theta}_{F_{N^*}}(\cdot)) &= N^* + 1 + \int_0^T \theta^2(t)\mu(dt) - \frac{T}{2}\sum_{j\in F_{N^*}} \theta_{j,T}^2 \\ IMSE(\hat{\theta}_{F_N}(\cdot)) &= N + 1 + \int_0^T \theta^2(t)\mu(dt) - \frac{T}{2}\sum_{j\in F_N} \theta_{j,T}^2. \end{aligned}$$

Therefore,

$$IMSE(\hat{\theta}_{F_{N^*}}(\cdot)) = IMSE(\hat{\theta}_{F_N}(\cdot)) + \frac{T}{2}\sum_{j_i\in F_N\setminus F_{N^*}} \theta_{j_i,T}^2 - (N - N^*).$$

Compared with $\hat{\theta}_{F_N}(\cdot)$, the estimator $\hat{\theta}_{F_{N^*}}(\cdot)$ has an increased ISB and a reduced IVAR. Note that $\frac{T}{2}\sum_{j_i \in F_N \setminus F_{N^*}} \theta^2_{j_i,T} \leq \sum_{j_i \in F_N \setminus F_{N^*}} 1 = N - N^*$ by the definition of K_{max} in *Step* 1. Therefore, we have $IMSE(\hat{\theta}_{F_{N^*}}(\cdot)) \leq IMSE(\hat{\theta}_{F_N}(\cdot))$.

The above result implies that we need to consider only the case $N \leq K_{max}+1$. Let $\Sigma_N = \{O_0, O_2, \cdots, O_{N-1}\}$, where $O_i's$ are as defined in *Step* 2. It then follows that $\hat{\theta}_{\Sigma_{K_{opt+1}}}(t) = \hat{\theta}_{T,K_{opt}}(t)$. We will show first that $\hat{\theta}_{\Sigma_N}(\cdot)$ has the smallest IMSE among all estimators of the form $\hat{\theta}_{F_N}(\cdot)$ and then prove that $IMSE(\hat{\theta}_{\Sigma_{K_{opt+1}}}(\cdot)) \leq IMSE(\hat{\theta}_{\Sigma_N}(\cdot))$, for every $N \leq K_{max}+1$.

To prove that $\hat{\theta}_{\Sigma_N}(\cdot)$ has the smallest IMSE among all estimators of the form $\hat{\theta}_{F_N}(\cdot)$ for any $N \leq K_{max}+1$, note that all $\hat{\theta}_{F_N}(\cdot)$s have the same IVAR equal to N and hence we need to find only the subset F_N with the smallest ISB. Note that $ISB(\hat{\theta}_{F_N}(\cdot)) = \int_0^T \theta^2(t)dt - \frac{T}{2}\sum_{j\in F_N}\theta^2_{j,T}$. Since, by definition, Σ_N contains the largest N terms of $\theta^2_{j,T}$, it follows that $\hat{\theta}_{\Sigma_N}(\cdot)$ has the smallest ISB and hence the smallest IMSE among all estimators of the form $\hat{\theta}_{F_N}(\cdot)$.

Next, we prove that $\hat{\theta}_{\Sigma_{K_{opt+1}}}(\cdot)$ has the smallest IMSE amongst all estimators of the form $\hat{\theta}_{\Sigma_N}(\cdot)$ for any $N \leq K_{max}$. Indeed, from the definition of K_{opt} in *Step* 3 and the form of IMSE in (4.21),

$$1 + \int_0^T \theta^2(t)\mu(dt) - \frac{T}{2}\theta^2_{O_0,T} \geq 2 + \int_0^T \theta^2(t)\mu(dt) - \frac{T}{2}\sum_{j=0}^{1}\theta^2_{O_j,T}$$

$$\geq \cdots \geq K_{opt} + 1 + \frac{T}{2}\sum_{j=0}^{K_{opt}}\theta^2_{O_j,T}.$$

This implies

$$IMSE(\hat{\theta}_{\Sigma_1}) \geq IMSE(\hat{\theta}_{\Sigma_2}) \geq \cdots \geq IMSE(\hat{\theta}_{\Sigma_{K_{opt+1}}}). \tag{4.32}$$

On the other hand, from the definition of K_{opt} in *Step* 3,

$$\begin{aligned} & K_{opt} + 1 + \int_0^T \theta^2(t)\mu(dt) - \frac{T}{2}\sum_{j=0}^{K_{opt}} \theta^2_{O_j,T} \\ \leq \; & K_{opt} + 2 + \int_0^T \theta^2(t)\mu(dt) - \frac{T}{2}\sum_{j=0}^{K_{opt+1}} \theta^2_{O_j,T} \\ \leq \; & \cdots \leq K_{max} + 1 + \int_0^T \theta^2(t)\mu(dt) - \frac{T}{2}\sum_{j=0}^{K_{max}} \theta^2_{O_j,T}, \end{aligned}$$

which implies

$$IMSE(\hat{\theta}_{\Sigma_{K_{opt+1}}}) \leq IMSE(\hat{\theta}_{\Sigma_{K_{opt+2}}}) \cdots \leq IMSE(\hat{\theta}_{\Sigma_{K_{max}+1}}). \tag{4.33}$$

Combining (4.32) and (4.33), we get $IMSE(\hat{\theta}_{\Sigma_{K_{opt+1}}}) \leq IMSE(\hat{\theta}_{\Sigma_N}), \forall\, N \leq K_{max}$. □

Chapter 5

Inference for Time-Inhomogeneous Mean-Reverting SDEs

5.1 Introduction

In this chapter, we study the following mean-reverting SDE:

$$dr(t) = a(\theta(t) - r(t))dt + \sigma(t, r(t))dW(t), \quad 0 \leq t \leq T, \tag{5.1}$$

where a is an unknown nuisance parameter and $\theta(t)$ is the unknown function of interest, which we call the level function. Suppose that a continuous realization of $\{r(t)\}$ is observed on the interval $[0, T]$. This implies that, for each t, the diffusion function value $\sigma(t, r(t))$ is known, as otherwise it can be estimated from the continuous realization of (5.1). We make the same assumptions about $\theta(t)$ and $\sigma(t, r(t))$ as in Chapter 4, i.e., $|\theta(t)| \leq \bar{M}$ and $0 < \bar{\epsilon} < \sigma(t, r(t)) < \bar{M}$ for some $\bar{\epsilon}$ and $\bar{M}$.

Using the same tools as in Section 3.4.1, we can represent $r(t)$ in the following form:

$$r(t) = e^{-at}r_0 + \int_0^t e^{-a(t-u)}\sigma(u, r(u))dW(u) + a\int_0^t e^{-a(t-u)}\theta(u)du. \quad (5.2)$$

In 2003, Fan et al. study the following model:

$$dX(t) = [\alpha_0(t) + \alpha_1(t)X(t)]dt + \beta_0(t)X(t)^{\beta_1(t)}dW(t).$$

The authors employ a local linear technique to estimate $\{\alpha_j\}$ and $\{\beta_j\}$ nonparametrically from discretely observed data. However, the asymptotic theory for their proposed method is still unknown. A similar model to (5.1) has been studied in Dehling et al. (2010). Their model has the following form:

$$dX(t) = (L(t) - \alpha X(t))dt + \sigma dW(t),$$

where σ is a known constant. The authors assume that the function $L_t = \sum_{i=1}^n \mu_i\phi_i(t)$ is periodic and parametric, and prove the asymptotic consistency and normality of the maximum likelihood estimator of μ_is as $T \to \infty$. In contrast, we assume that $\{\theta(t)\}$ is non-parametric without periodicity constraints. Our objective is to estimate $M_{T,K}(\theta)(t)$, the projection of the level function $\{\theta(t)\}$ onto a finite dimensional space, from a continuously observed single trajectory.

The methodology in this chapter is similar to that in Chapter 4, but the inference problem is more challenging. The main reason is that, unlike the results in Chapter 4, we do not have an explicit distribution of the maximum likelihood estimators. Therefore, the proof of asymptotic results and the construction of a confidence interval cannot be obtained with the same analytical tools developed in Chapter 4.

In finite sample studies, we have found that the estimation error for $M_{T,K}(\theta)(t)$ is much larger than that obtained in Chapter 4. In fact, the estimation error for $M_{T,K}(\theta)$ depends on the estimation accuracy for a, which in turn depends on how well $M_{T,K}(\theta)(t)$ approximates the true level function $\theta(t)$. This "aliasing" problem

between estimation of a and estimation of $M_{T,K}(\theta)(t)$ is the main additional challenge when compared with the inference problem in Chapter 4. It turns out that the accuracy of the mean-reverting speed estimator is crucial in determining the estimation error for $M_{T,K}(\theta)(t)$ for fixed t. Our finite sample simulation studies show that the mean-reverting speed can suffer from serious upward bias.

In the econometrics literature, it is known that estimation of the mean-reverting speed parameter in the Vasicek model, which is a special case of (5.1) with both $\theta(t)$ and $\sigma(t, r(t))$ constants, is quite challenging (Ball and Torous, 1996; Yu and Philips, 2001, 2005; Tang and Chen, 2009; Yu, 2011). This estimation difficulty arises because standard estimation methods often yield biased estimators of this parameter. Tang and Chen (2009) conduct a simulation study using the following parameter values: $a = 0.215, \sigma = 0.0224$ and $\theta(t) = 0.089$. The authors conduct 5000 simulations of 10 years of monthly data and find that the relative bias of the mean-reverting speed parameter can be over 200%. Given the significant estimation error in the mean-reverting speed parameter, estimating a projection of the level function is more difficult than estimating the projection of the drift function of a Brownian motion. In this chapter, we propose a basic parametric bootstrap method, similar to the one used in Tang and Chen (2009), to correct the bias for the mean-reverting speed estimator. The results are encouraging.

Our contributions to methods of statistical inference on $M_{T,K}(\theta)(t)$ for fixed t can be summarized as follows:

P1. We derive closed-form maximum likelihood estimator for both a and $M_{T,K}(\theta)(t)$.

P2. We prove that as long as the dimension of the projection space grows at a controlled speed with T, both the estimators of a and $M_{T,K}(\theta)(t)$ are weakly consistent.

P3. Unlike the results in Chapter 4, we do not have an explicit distribution of

the maximum likelihood estimator of a and $M_{T,K}(\theta)(t)$. We have derived the asymptotic distribution of $M_{T,K}(\theta)(t)$. For daily data of length over 50 years, we rely on the asymptotic results to determine the dimension of parameter space and construct an approximate point-wise confidence interval for $M_{T,K}(\theta)(t)$.

P4. For daily data of length shorter than 50 years, we propose a basic parametric bootstrap method to determine the dimension of the parameter space and construct a confidence interval for $M_{T,K}(\theta)(t)$.

The layout of the rest of this chapter is as follows:

- Section 5.2 derives the restricted maximum likelihood estimator for both the mean-reverting speed parameter and the projection of the level function onto a finite dimensional space.
- Section 5.3 proves the asymptotic consistency and normality of the maximum likelihood estimator when the dimension of the projection space increases with the data length at a controlled speed.
- Section 5.4 proposes two methods for determining the dimension of the parameter space and constructing a confidence interval for the projection of the level function. Section 5.5 presents simulation results for the proposed methods, while Section 5.6 applies the proposed methodology to an interest rate data set.
- Section 5.7 draws concluding remarks for this chapter.

5.2 The Maximum Likelihood Estimator

The method we use to define the maximum likelihood estimator in this section is the same as in Section 4.3.2, except that more work is needed to obtain analytical

forms of the estimator for a and $M_{T,K}(\theta)(t)$.

Assume that $\theta(\cdot)$ and $\sigma(\cdot)$ satisfy the Lipschitz continuity and linear growth conditions that ensure the existence of a unique strong solution to (5.1). Let $\mu(dt)$ and $\{p_{i,T}, i = 0, 1, \cdots\}$ be the same as in Chapter 4. Since $\{\theta(t)\} \in L^2([0,T], \mu(dt))$, we can write

$$\theta(t) = M_{T,K}(\theta)(t) + M^{\perp}_{T,K}(\theta)(t) = \vec{p}_{T,K}(t)'\Phi_{T,K}(\theta) + M^{\perp}_{T,K}(\theta)(t), \quad 0 \le t \le T.$$

Since $\{p_{i,T}, i \ge 0\}$ are linearly independent functions on $[0,T]$, we have a unique representation of $\Phi_{T,K}(\theta) = (\theta_{0,T}, \theta_{1,T}, \cdots, \theta_{K,T})'$. Let P_r^T be the probability measure generated by the process $\{r(t) : 0 \le t \le T\}$ on the space $(C[0,T], \mathcal{B}_T)$. Let P_η^T be the probability measure induced by the strong solution to the following SDE:

$$d\eta(t) = \sigma(t, \eta(t))dW(t), \quad 0 \le t \le T.$$

Since $\{r(t)\}$, as a solution to (5.1), is a continuous semi-martingale, its paths are bounded on the compact interval $[0,T]$. By our assumption in Section 5.1, $|\theta(t)| \le \bar{M}$ and $0 < \bar{\epsilon} < \sigma(t, r(t)) < \bar{M}$. Then we have the following:

$$P(\omega \in \Omega : |\frac{a(\theta(t) - r(t))}{\sigma(t, r(t))}| < \infty) = 1, \quad \forall \quad 0 \le t \le T,$$
$$P(\omega \in \Omega : \int_0^T \frac{a^2(\theta(t) - r(t))^2}{\sigma^2(t, r(t))} dt < \infty) = 1.$$

Therefore, we have $P_r^T \ll p_\eta^T$ (Lipster and Shiryaev, 1974) and

$$\begin{aligned}
\frac{dP_r^T}{dP_\eta^T} &= \exp\{\int_0^T \frac{a(\theta(t) - r(t))}{\sigma^2(t, r(t))} dr(t) - \frac{1}{2}\int_0^T \frac{a^2(\theta(t) - r(t))^2}{\sigma^2(t, r(t))} dt\} \\
&= \exp\{\int_0^T \frac{a(\vec{p}'_{T,K}(t)\Phi_{T,K}(\theta) + M^{\perp}_{T,K}(\theta)(t) - r(t))}{\sigma^2(t, r(t))} dr(t) \\
&\quad -\frac{1}{2}\int_0^T \frac{a^2(\vec{p}'_{T,K}(t)\Phi_{T,K}(\theta) + M^{\perp}_{T,K}(\theta)(t) - r(t))^2}{\sigma^2(t, r(t))} dt\}.
\end{aligned}$$

By treating $M^{\perp}_{T,K}(\theta)(t)$ as a nuisance parameter, we can derive the maximum likelihood estimator of a and $\Phi_{T,K}(\theta)$ by maximizing the log-likelihood function. Let

us denote the log-likelihood function by l. Then the score functions for a and $\theta_{m,T}, 0 \leq m \leq K$ can be represented as follows:

$$\begin{aligned}
\frac{\partial l}{\partial a} &= \int_0^T \frac{(\vec{p'}_{T,K}(t)\Phi_{T,K}(\theta) + M^{\perp}_{T,K}(\theta)(t) - r(t))}{\sigma(t,r(t))^2} dr(t) \\
&\quad - \int_0^T \frac{a(\vec{p'}_{T,K}(t)\Phi_{T,K}(\theta) + M^{\perp}_{T,K}(\theta)(t) - r(t))^2}{\sigma(t,r(t))^2} dt, \\
\frac{\partial l}{\partial \theta_{m,T}} &= \int_0^T \frac{a p_{m,T}(t)}{\sigma(t,r(t))^2} dr(t) - \int_0^T \frac{a^2(\vec{p'}_{T,K}(t)\Phi_{T,K}(\theta) + M^{\perp}_{T,K}(\theta)(t) - r(t))p_{m,T}(t)}{\sigma(t,r(t))^2} dt \\
&= \int_0^T \frac{a p_{m,T}(t)}{\sigma(t,r(t))^2} dr(t) - \int_0^T \frac{a^2 \vec{p'}_{T,K}(t)\Phi_{T,K}(\theta) p_{m,T}(t)}{\sigma(t,r(t))^2} dt + \int_0^T \frac{a^2 r(t) p_{m,T}(t)}{\sigma(t,r(t))^2} dt,
\end{aligned}$$

where the last equality holds because $M^{\perp}_{T,K}(\theta)(t)$ and $V_{T,K}$ are orthogonal in $L^2([0,T], \mu(dt))$. As roots of the above score functions, the maximum likelihood estimator satisfy the following equations:

$$\begin{aligned}
\tilde{a}_{T,K} &= \frac{\int_0^T \frac{(\vec{p'}_{T,K}(t)\hat{\Phi}_{T,K}(\theta) + M^{\perp}_{T,K}(\theta)(t) - r(t))}{\sigma(t,r(t))^2} dr(t)}{\int_0^T \frac{(\vec{p'}_{T,K}(t)\hat{\Phi}_{T,K}(\theta) + M^{\perp}_{T,K}(\theta)(t) - r(t))^2}{\sigma(t,r(t))^2} dt} \\
\tilde{\Phi}_{T,K}(\theta) &= A^{-1}_{T,K} \int_0^T \frac{\vec{p}_{T,K}(t) r(t)}{\sigma(t,r(t))^2} dt + \frac{1}{\tilde{a}_{T,K}} A^{-1}_{T,K} \int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2} dr(t), \quad (5.3)
\end{aligned}$$

where $A_{T,K}$ is the same as the one introduced in Chapter 4. Since the expression of $\tilde{a}_{T,K}$ involves the unknown nuisance parameter $M^{\perp}_{T,K}(\theta)(t)$, we propose the following modified maximum likelihood estimator of a and $\Phi_{T,K}(\theta)$:

$$\begin{aligned}
\hat{a}_{T,K} &= \frac{\int_0^T \frac{(\vec{p'}_{T,K}(t)\hat{\Phi}_{T,K}(\theta) - r(t))}{\sigma(t,r(t))^2} dr(t)}{\int_0^T \frac{(\vec{p'}_{T,K}(t)\hat{\Phi}_{T,K}(\theta) - r(t))^2}{\sigma(t,r(t))^2} dt} \quad (5.4) \\
\hat{\Phi}_{T,K}(\theta) &= A^{-1}_{T,K} \int_0^T \frac{\vec{p}_{T,K}(t) r(t)}{\sigma(t,r(t))^2} dt + \frac{1}{\hat{a}_{T,K}} A^{-1}_{T,K} \int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2} dr(t). \quad (5.5)
\end{aligned}$$

The idea is that if $\theta(\cdot) \in V_{T,K}$, then $M^{\perp}_{T,K}(\theta)(t) = 0$ for each $t \in [0,T]$ and $\hat{a}_{T,K} = \tilde{a}_{T,K}$. Otherwise, by selecting a large K, we can assume that $M^{\perp}_{T,K}(\theta)(t)$ is

small enough. We will be referring to these modified maximum likelihood estimator as maximum likelihood estimator.

Note that the maximum likelihood estimator defined by (5.4) and (5.5) are not in a closed form, as $\hat{a}_{T,K}$ and $\hat{\Phi}_{T,K}(\theta)$ appear in both equations. The following theorem gives the closed-form expressions of maximum likelihood estimator $\hat{a}_{T,K}$ and $\hat{\Phi}_{T,K}(\theta)$. The technical proof is presented in Appendix 5.8.

Theorem 5.1. *Let $\hat{a}_{T,K}$ and $\hat{\Phi}_{T,K}(\theta)$ be solutions to the system of equations given in (5.4) and (5.5). Then we have the following representation of $\hat{a}_{T,K}$:*

$$\hat{a}_{T,K} = \frac{-\int_0^T \frac{M_{T,K}^{\perp}(r)(t)}{\sigma(t,r(t))^2} dr(t)}{\int_0^T \frac{[M_{T,K}^{\perp}(r)(t)]^2}{\sigma(t,r(t))^2} dt}. \tag{5.6}$$

We can also obtain a closed-form expression for $\hat{\Phi}_{T,K}(\theta)$ by replacing $\hat{a}_{T,K}$ in (5.5) with (5.6).

We would like to emphasize that the closed-form derived in (5.6) and the form (5.4) are different, although they look similar. The difference is that while (5.4) depends on the estimator $\hat{\Phi}_{T,K}(\theta)$, (5.6) does not.

In the following, we derive convenient forms of $\hat{\Phi}_{T,K}(\theta)$ defined in (5.5), so that subsequent presentation of the confidence intervals and hypothesis-testing procedures is easier. The idea is to separate the term $\Phi_{T,K}(\theta)$ from the expression of $\hat{\Phi}_{T,K}(\theta)$. We use the fact that, by definition, $\Phi_{T,K}(\theta) = A_{T,K}^{-1} \int_0^T \frac{\vec{p}_{T,K}(t)\theta(t)}{\sigma(t,r(t))^2} dt$. With

$dr(t)$ replaced by the right-hand side of (5.1), we can rewrite (5.5):

$$\begin{aligned}
&\hat{\Phi}_{T,K}(\theta) \\
= \;& A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)(\theta(t)+r(t)-\theta(t))}{\sigma(t,r(t))^2}dt \\
&+\frac{1}{\hat{a}_{T,K}}A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}\Big(a\left(\theta(t)-r(t)\right)dt+\sigma(t,r(t))dW(t)\Big) \\
= \;& \Phi_{T,K}(\theta)+A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)(r(t)-\theta(t))}{\sigma(t,r(t))^2}dt \\
&+\frac{a}{\hat{a}_{T,K}}A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}(\theta(t)-r(t))dt+\frac{1}{\hat{a}_{T,K}}A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))}dW(t) \\
= \;& \Phi_{T,K}(\theta)+\Phi_{T,K}(r-\theta)+\frac{a}{\hat{a}_{T,K}}\Phi_{T,K}(\theta-r)+\frac{1}{\hat{a}_{T,K}}A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))}dW(t) \\
= \;& \Phi_{T,K}(\theta)+\frac{a-\hat{a}_{T,K}}{\hat{a}_{T,K}}\Phi_{T,K}(\theta-r)+\frac{1}{\hat{a}_{T,K}}A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))}dW(t).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\hat{a}_{T,K}(\hat{\Phi}_{T,K}(\theta)-\Phi_{T,K}(\theta))-(a-\hat{a}_{T,K})\Phi_{T,K}(\theta-r) \\
= \;& A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))}dW(t). \qquad (5.7)
\end{aligned}$$

Define $\hat{M}_{T,K}(\theta)(t) \triangleq \vec{p}_{T,K}(t)'\hat{\Phi}_{T,K}(\theta)$. Multiplying both sides of (5.7) by $\vec{p'}_{T,K}(t)$, we get

$$\begin{aligned}
&\hat{a}_{T,K}[\hat{M}_{T,K}(\theta)(t)-M_{T,K}(\theta)(t)]-(a-\hat{a}_{T,K})M_{T,K}(\theta-r)(t) \\
= \;& \vec{p'}_{T,K}(t)A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))}dW(t). \qquad (5.8)
\end{aligned}$$

The left-hand side of expression (5.8) involves the parameters a and $M_{T,K}(\theta)(t)$. When the distribution of the right-hand side of (5.8) is known, we can construct a joint confidence region for a and $M_{T,K}(\theta)(t)$. In order to obtain an explicit distribution of $\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))}dW(t)$, we consider the following two cases:

Case 1: $\vec{p}_{T,K}(t) = \sigma(t,r(t))\vec{q}_{T,K}(t)$.

Case 2: $\vec{p}_{T,K}(t) = \vec{q}_{T,K}(t)$ and $\sigma(t, r(t)) = \sigma(t)$; that is, the diffusion term is modeled as a function of time t only.

The two cases correspond to different sets of assumptions about $\sigma(\cdot, \cdot)$ and the selection of basis functions $\vec{p}_{T,K}(t)$ we are going to use. The first case is introduced because of the mathematical ease of deriving limiting properties of the proposed estimators. The second case is introduced because we find in our empirical analysis that the procedures developed in this case lead to more robust estimation results.

In the following, we present a result on the confidence interval for the projected mean-reversion level function.

Theorem 5.2. *In both Case* 1 *and* 2*, we have*

$$\hat{a}_{T,K}(\hat{\Phi}_{T,K}(\theta) - \Phi_{T,K}(r)) + a(\Phi_{T,K}(r) - \Phi_{T,K}(\theta)) \sim MVN(0, A_{T,K}^{-1}), \tag{5.9}$$

and a 95% *confidence interval for* $M_{T,K}(\theta)(t)$ *is*

$$\left(M_{T,K}(r)(t) + \frac{\hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(r)(t)) - z_{0.975}\sqrt{\vec{p}'_{T,K}(t) A_{T,K}^{-1} \vec{p}_{T,K}(t)}}{a}, \right.$$
$$\left. M_{T,K}(r)(t) + \frac{\hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(r)(t)) - z_{0.025}\sqrt{\vec{p}'_{T,K}(t) A_{T,K}^{-1} \vec{p}_{T,K}(t)}}{a} \right). \tag{5.10}$$

Proof. We proceed to prove the theorem for each case separately. In Case 1, since $\vec{p}_{T,K}(t) = \sigma(t, r(t))\vec{q}_{T,K}(t)$, we have

$$\begin{aligned} \int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t, r(t))} dW(t) = \int_0^T \vec{q}_{T,K}(t) dW(t) &\sim MVN(0, \int_0^T \vec{q}_{T,K}(t)\vec{q}_{T,K}(t)' dt) \\ &= MVN(0, \frac{T}{2} I_{(K+1)\times(K+1)}), \end{aligned}$$

where the last equation follows from the definition of $\vec{q}_{T,K}(t)$ in Section 4.3.1. Under this assumption, the representation (5.7) implies that

$$\hat{a}_{T,K}(\hat{\Phi}_{T,K}(\theta) - \Phi_{T,K}(\theta)) - (a - \hat{a}_{T,K})\Phi_{T,K}(\theta - r) \sim MVN(0, A_{T,K}^{-1}), \tag{5.11}$$

and the representation (5.8) implies that

$$\frac{\hat{a}_{T,K}[\hat{M}_{T,K}(\theta)(t) - M_{T,K}(\theta)(t)] - (a - \hat{a}_{T,K})M_{T,K}(\theta - r)(t)}{\sigma(t, r(t))}$$
$$\sim N(0, \vec{q}_{T,K}(t)' A_{T,K}^{-1} \vec{q}_{T,K}(t)), \tag{5.12}$$

noting that $A_{T,K} = \frac{T}{2} I_{(K+1)\times(K+1)}$ in this case. Below we derive alternative forms of (5.11) and (5.12). These forms will be used to construct a confidence interval for $M_{T,K}(\theta)(t)$. We reorganize the left-hand side of (5.11):

$$\begin{aligned} & \hat{a}_{T,K}[\hat{\Phi}_{T,K}(\theta) - \Phi_{T,K}(\theta)] + (\hat{a}_{T,K} - a)\Phi_{T,K}(\theta - r) \\ = \; & \hat{a}_{T,K}\hat{\Phi}_{T,K}(\theta) - \hat{a}_{T,K}\Phi_{T,K}(\theta) + \hat{a}_{T,K}\Phi_{T,K}(\theta) - a\Phi_{T,K}(\theta) \\ & -\hat{a}_{T,K}\Phi_{T,K}(r) + a\Phi_{T,K}(r) \\ = \; & \hat{a}_{T,K}\hat{\Phi}_{T,K}(\theta) - \hat{a}_{T,K}\Phi_{T,K}(r) + a\Phi_{T,K}(r) - a\Phi_{T,K}(\theta). \end{aligned}$$

Then (5.11) implies the following

$$\hat{a}_{T,K}(\hat{\Phi}_{T,K}(\theta) - \Phi_{T,K}(r)) + a(\Phi_{T,K}(r) - \Phi_{T,K}(\theta)) \sim MVN(0, A_{T,K}^{-1}).$$

Similarly, we reorganize the left-hand side of (5.12) to obtain

$$\begin{aligned} & \hat{a}_{T,K}[\hat{M}_{T,K}(\theta)(t) - M_{T,K}(\theta)(t)] + (\hat{a}_{T,K} - a)M_{T,K}(\theta - r)(t) \\ = \; & \hat{a}_{T,K}\hat{M}_{T,K}(\theta)(t) - \hat{a}_{T,K}M_{T,K}(r)(t) + aM_{T,K}(r)(t) - aM_{T,K}(\theta)(t). \end{aligned}$$

Then (5.12) implies the following

$$\frac{\hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(r)(t)) + a(M_{T,K}(r)(t) - M_{T,K}(\theta)(t))}{\sigma(t, r(t))}$$
$$\sim N(0, \vec{q}'_{T,K}(t) A_{T,K}^{-1} \vec{q}_{T,K}(t)).$$

From the above equation, a 95% confidence interval for $M_{T,K}(\theta)(t)$ is given by

$$\left(M_{T,K}(r)(t) + \frac{\hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(r)(t))}{a} - \frac{z_{0.975}\sigma(t, r(t))\sqrt{\vec{q}'_{T,K}(t) A_{T,K}^{-1} \vec{q}_{T,K}(t)}}{a}, \right.$$

$$\left. M_{T,K}(r)(t) + \frac{\hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(r)(t))}{a} - \frac{z_{0.025}\sigma(t, r(t))\sqrt{\vec{q}'_{T,K}(t) A_{T,K}^{-1} \vec{q}_{T,K}(t)}}{a} \right).$$

Since $\vec{p}_{T,K}(t) = \sigma(t, r(t))\vec{q}_{T,K}(t)$ in this case, the above confidence interval can be rewritten as

$$\left(M_{T,K}(r)(t) + \frac{\hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(r)(t)) - z_{0.975}\sqrt{\vec{p}'_{T,K}(t)A^{-1}_{T,K}\vec{p}_{T,K}(t)}}{a}, \right.$$
$$\left. M_{T,K}(r)(t) + \frac{\hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(r)(t)) - z_{0.025}\sqrt{\vec{p}'_{T,K}(t)A^{-1}_{T,K}\vec{p}_{T,K}(t)}}{a} \right).$$

In case 2, since $\vec{p}_{T,K}(t) = \vec{q}_{T,K}(t)$ and $\sigma(t, r(t)) = \sigma(t)$, we have an explicit distribution of $\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))}dW(t)$ as follows:

$$\begin{aligned} \int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t)}dW(t) &\sim MVN(0, \int_0^T \frac{\vec{p}_{T,K}(t)\vec{p}_{T,K}(t)'}{\sigma^2(t)})dt \\ &= MVN(0, A^{-1}_{T,K}). \end{aligned}$$

The representation (5.7) implies that

$$\hat{a}_{T,K}(\hat{\Phi}_{T,K}(\theta) - \Phi_{T,K}(\theta)) - (a - \hat{a}_{T,K})\Phi_{T,K}(\theta - r) \sim MVN(0, A^{-1}_{T,K}), \quad (5.13)$$

and the representation (5.8) implies that

$$\begin{aligned} &\hat{a}_{T,K}[\hat{M}_{T,K}(\theta)(t) - M_{T,K}(\theta)(t)] - (a - \hat{a}_{T,K})M_{T,K}(\theta - r)(t) \\ &\sim N(0, \vec{q}_{T,K}(t)'A^{-1}_{T,K}\vec{q}_{T,K}(t)). \end{aligned} \quad (5.14)$$

Similar to Case 1, we have the following alternative forms of (5.13) and (5.14):

$$\hat{a}_{T,K}(\hat{\Phi}_{T,K}(\theta) - \Phi_{T,K}(r)) + a(\Phi_{T,K}(r) - \Phi_{T,K}(\theta)) \sim MVN(0, A^{-1}_{T,K})$$

and

$$\begin{aligned} &\hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(r)(t)) + a(M_{T,K}(r)(t) - M_{T,K}(\theta)(t)) \\ &\sim N(0, \vec{q}'_{T,K}(t)A^{-1}_{T,K}\vec{q}_{T,K}(t)). \end{aligned}$$

From the above equation, a 95% confidence interval for $M_{T,K}(\theta)(t)$ is then given by

$$\left(M_{T,K}(r)(t) + \frac{\hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(r)(t))}{a} - \frac{z_{0.975}\sqrt{\vec{q}'_{T,K}(t)A^{-1}_{T,K}\vec{q}_{T,K}(t)}}{a}, \right.$$

$$\left. M_{T,K}(r)(t) + \frac{\hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(r)(t))}{a} - \frac{z_{0.025}\sqrt{\vec{q}'_{T,K}(t)A^{-1}_{T,K}\vec{q}_{T,K}(t)}}{a} \right).$$

Since $\vec{p}_{T,K}(t) = \vec{q}_{T,K}(t)$ in this case, the above confidence interval can be rewritten as

$$\left(M_{T,K}(r)(t) + \frac{\hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(r)(t)) - z_{0.975}\sqrt{\vec{p}'_{T,K}(t)A^{-1}_{T,K}\vec{p}_{T,K}(t)}}{a}, \right.$$

$$\left. M_{T,K}(r)(t) + \frac{\hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - M_{T,K}(r)(t)) - z_{0.025}\sqrt{\vec{p}'_{T,K}(t)A^{-1}_{T,K}\vec{p}_{T,K}(t)}}{a} \right).$$

□

Note that the derived confidence interval (5.10) contains the unknown mean-reverting speed parameter 'a'. Because of this, in Section 5.4.2 we first estimate the range of 'a' and then derive an approximate 95% confidence interval of $M_{T,K}(\theta)(t)$ based on (5.10). Another important observation that we can make about this confidence interval is that its width is proportional to $1/a$. Therefore, a low mean-reverting speed parameter implies a wider confidence interval for $M_{T,K_T}(\theta)(t)$.

The definition of the basis function we use in Case 1 is convenient, since then $A_{T,K} = \frac{T}{2}I_{(K+1)\times(K+1)}$, which simplifies both (5.9) and (5.10). Therefore, finding the limit is feasible. The asymptotic results in Section 5.3 hold only in this case. In Case 2, we are unable to derive asymptotic properties, such as weak consistency of $\hat{M}_{T,K}(\theta)(t)$. However, we propose a method to determine the dimension of the parameter space and construct a confidence interval for $M_{T,K}(\theta)(t)$. The proposed method works well even for small examples, i.e, when the data length T is moderate.

In addition, our empirical analysis shows that the proposed method in Case 2 seems to be more robust to deviations from model assumptions.

5.3 Asymptotic Results: a Sieve-type Approach

In this section, we allow the dimension parameter K to increase as the time length T increases. Therefore, we introduce the notation K_T to indicate dependence of K on T. In the following, we prove that the maximum likelihood estimator of a and $M_{T,K}(\theta)(t)$ are weakly consistent and asymptotically normally distributed. The conditions required for these results to hold include that K_T increases with T at a controlled speed. In this section, we consider only Case 1, i.e., $\vec{p}_{T,K_T} = \sigma(t, r(t))\vec{q}_{T,K_T}$. It follows from the definition of $\vec{q}_{T,K_T}$ in Section 4.3.1 that $\vec{p}_{T,K_T}$ is an orthogonal vector in $L^2([0,T], \mu(dt))$. In this case, we have the simplification $A_{T,K_T} = \frac{T}{2} I_{(K_T+1)\times(K_T+1)}$, which facilitates our proofs. We should mention that proofs in this section are more complicated than the ones in Section 4.3.3, due to the aliasing problem of estimating a and $M_{T,K}(\theta)(t)$ at the same time. All the technical proofs are presented in Appendix 5.8.

Asymptotic Consistency

The following theorem states that $\hat{a}_{T,K_T}$ converges in probability to a under certain technical conditions. In particular, the dimension parameter K_T cannot increase too fast with T.

Theorem 5.3. *The maximum likelihood estimator (5.6) for the parameter a in model (5.1) is weakly consistent provided that the following conditions hold.*

($\mathcal{A}$1) There exists $\delta > 0$ such that, with probability one, $\dfrac{\int_0^T [M_{T,K_T}(r)(t)]^2 \mu(dt)}{\int_0^T r(t)^2 \mu(dt)} <$

$1-\delta$ *for all* T *sufficiently large, i.e.,*

$$P(\limsup_{T\to\infty} \frac{\int_0^T [M_{T,K_T}(r)(t)]^2 \mu(dt)}{\int_0^T r(t)^2 \mu(dt)} < 1-\delta) = 1.$$

($\mathcal{A}2$*)* $\lim_{T\to\infty} \dfrac{\int_0^T [M^{\perp}_{T,K_T}(\theta)(t)]^2 \mu(dt)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2 \mu(dt)} = 0,$ *in probability.*

($\mathcal{A}3$*)* $P(\lim_{T\to\infty} \dfrac{K_T+1}{\int_0^T r(t)^2\mu(dt)} = 0) = 1.$

The proof of the above theorem uses the following decomposition

$$\hat{a}_{T,K_T} = a - a\frac{\int_0^T M^{\perp}_{T,K_T}(r)(t) M^{\perp}_{T,K_T}(\theta)(t)\mu(dt)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2\mu(dt)} - \frac{\int_0^T \frac{M^{\perp}_{T,K_T}(r)(t)}{\sigma(t,r(t))} dW(t)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2\mu(dt)}, \quad (5.15)$$

which is shown in Appendix 5.8. The second term on the right hand side of the above equation converges to zero in probability under condition $\mathcal{A}2$; the third term converges to zero in probability under conditions $\mathcal{A}1$ and $\mathcal{A}3$.

Condition $\mathcal{A}1$ requires K_T not to increase too fast with T, so that $\int_0^T [M_{T,K_T}(r)(t)]^2\mu(dt)$ is strictly less than $\int_0^T r(t)^2\mu(dt)$. When K_T is large enough, however, $\int_0^T [M_{T,K_T}(r)(t)]^2\mu(dt)$ can be arbitrarily close to $\int_0^T r(t)^2\mu(dt)$. Condition $\mathcal{A}2$ imposes certain constraints on $\{\theta(t), t \geq 0\}$. It is clearly satisfied when $\theta(t)$ is a polynomial, since then $M^{\perp}_{T,K_T}(\theta)(t) = 0$ when K_T is larger than the order of the polynomial. Condition $\mathcal{A}3$ is another condition on the rate of K_T. In practice, conditions $\mathcal{A}1$ and $\mathcal{A}3$ are easy to check. Condition $\mathcal{A}2$ is more difficult to verify for a general function θ.

Condition $\mathcal{A}3$ can be replaced by the following two weaker conditions.

($\mathcal{A}4$*)* $P(\lim_{T\to\infty} \int_0^T r(t)^2\mu(dt) = \infty) = 1.$

($\mathcal{A}5$*)* $\lim_{T\to\infty} \dfrac{K_T+1}{\int_0^T r(t)^2\mu(dt)} = 0$ *in probability.*

In fact, the proof of Theorem 5.3 holds exactly the same with condition $\mathcal{A}3$ replaced by conditions $\mathcal{A}4$ and $\mathcal{A}5$. Condition $\mathcal{A}4$ is needed for applying the following Lemma 5.1, and condition $\mathcal{A}5$ is used for proof of (5.33) in Appendix 5.8.

For the proof of Theorem 5.3, the following result from Lipster and Shiryaev will be used:

Lemma 5.1. *Let the Brownian motion, $W = (W_t, \mathcal{F}_t), t \geq 0$, be given on a probability space and $f = (f_t, \mathcal{F}_t), t \geq 0$ be a stochastic process such that*

$$P(\int_0^T f_t^2 dt < \infty) = 1, \quad 0 < T < \infty;$$
$$P(\int_0^\infty f_t^2 dt = \infty) = 1.$$

Then the random process $z = (z_s, \mathcal{G}_s), s \geq 0$, with $z_s = \int_0^{\tau_s} f_t dW(t), \mathcal{G}_s = \mathcal{F}_{\tau_s}$, where $\tau_s = inf(t : \int_0^t f_u^2 du > s)$ is a Brownian motion and with probability one

$$\lim_{t\to\infty} \frac{\int_0^t f_u dW(u)}{\int_0^t f_u^2 du} = 0.$$

The second part of this Lemma can be interpreted as a version of the strong law of large numbers for continuous martingales.

In the next result below, we show that, for fixed t, $\hat{M}_{T,K}(\theta)(t)$ is a weakly consistent estimator of $M_{T,K}(t)$. Before we state the theorem, we need the following two lemmas.

Lemma 5.2. *Suppose that $f(t), \quad t \geq 0$ is a square integrable function in $L^2([0,T], \mu(dt))$. Then for each $t \in [0,T]$, we have*

$$|M_{T,K}(f)(t)| \leq \frac{\sigma(t, r(t))}{\bar{\epsilon}} \sqrt{\frac{2\int_0^T f^2(t)dt}{T} \sum_{j=0}^{K_T} q_{j,T}^2(t)},$$

where $\bar{\epsilon}$ is a lower bound for $\sigma(t, r(t))$, introduced at the beginning of this chapter.

This result provides bounds on the projection $M_{T,K}(f)(t)$ in terms of the L^2 norm of $f(t)$ on $[0,T]$.

Lemma 5.3. *Assume that $\{r(t)\}$ is the solution to (5.1). Then*

$$\sup_{T>0} E\left(\frac{1}{T}\int_0^T (r(t)-\theta(t))^2 dt\right) \leq 4\bar{M} + 2|r_0| + \bar{M}^2, \tag{5.16}$$

where $\bar{M}$ is an upper bound for both $\sigma(t,r(t))$ and $\theta(t)$, introduced at the beginning of this chapter.

One immediate consequence of the above lemma is that $\frac{1}{T}\int_0^T (r(t)-\theta(t))^2 dt$ is uniformly bounded in probability for all $T>0$. In fact, for any $\epsilon>0$, the Chebyshev's inequality and (5.16) imply that

$$\begin{aligned} & P\left(\frac{1}{T}\int_0^T (r(t)-\theta(t))^2 dt > \sqrt{\frac{\max(4\bar{M}+2|r_0|,\bar{M}^2)}{\epsilon}}\right) \\ \leq & \frac{E[\frac{1}{T}\int_0^T (r(t)-\theta(t))^2 dt]}{\frac{\max(4\bar{M}+2|r_0|,\bar{M}^2)}{\epsilon}} \\ \leq & \frac{\max(4\bar{M}+2|r_0|,\bar{M}^2)}{\frac{\max(4\bar{M}+2|r_0|,\bar{M}^2)}{\epsilon}} = \epsilon, \end{aligned}$$

for all $T>0$. Therefore, $\frac{1}{T}\int_0^T (r(t)-\theta(t))^2 dt$ is uniformly bounded in probability for $T>0$.

Theorem 5.4. *Assume that $\{\vec{R}(t) : -1 \leq t \leq 1\}$ are either normalized Legendre polynomials or trigonometric polynomials. Assume also that the conditions $\mathcal{C}1$ and $\mathcal{C}2$ in Theorem 4.2 are satisfied. Then $\hat{M}_{T,K_T}(\theta)(t)$ is a weakly consistent estimator of $M_{T,K_T}(\theta)(t)$ if the following condition is satisfied:*

($\mathcal{A}6$)

$$\lim_{T\to\infty} (\hat{a}_{T,K_T} - a)\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)} = 0 \text{ in probability.}$$

Conditions $\mathcal{C}1$ and $\mathcal{C}2$ require that K_T does not increase too fast with T. Condition $\mathcal{A}6$ requires that $\hat{a}_{T,K_T}$ converges to a faster than the rate of $\frac{1}{\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)}}$.

Note that in Theorem 5.3 we have proven, under conditions $\mathcal{A}1$, $\mathcal{A}2$ and $\mathcal{A}3$, that $\hat{a}_{T,K_T}$ converges to a in probability. The proof used the decomposition in (5.15). By examining the proof of Theorem 5.3, we can prove the following result.

Corollary 5.1. *Condition $\mathcal{A}6$ is satisfied if condition $\mathcal{A}1$ and the following conditions hold.*

$$(\mathcal{A}2+)\ \lim_{T\to\infty} \frac{\sum_{i=0}^{K_T} q_{i,T}^2(t) \int_0^T \left(M_{T,K_T}^{\perp}(\theta)(t)\right)^2 \mu(dt)}{\int_0^T \left(M_{T,K_T}^{\perp}(r)(t)\right)^2 \mu(dt)} = 0, \ \textit{in probability.}$$

$$(\mathcal{A}3+)\ P\left(\lim_{T\to\infty} \frac{(K_T+1)\sum_{i=0}^{K_T} q_{i,T}^2(t)}{\int_0^T r(t)^2 \mu(dt)} = 0\right) = 1.$$

$$(\mathcal{A}7)\ \lim_{T\to\infty} \frac{\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)} \int_0^T \frac{r(t)}{\sigma(t,r(t))} dW(t)}{\int_0^T r^2(t)\mu(dt)} = 0, \ \textit{in probability.}$$

Since the proof for the above result is similar to the one for Theorem 5.3, we omit the details. Condition $\mathcal{A}2+$ is a stronger version of $\mathcal{A}2$ to ensure the faster convergence of the second term on the right-hand side of (5.15) to zero. Conditions $\mathcal{A}1$, $\mathcal{A}3+$ and $\mathcal{A}7$ jointly ensure a faster convergence of the third term on the right-hand side of (5.15) to zero. In particular, condition $\mathcal{A}7$ is required to prove that the first term in (5.28) (in the proof of Theorem 5.3) converges to zero in probability. In practice, condition $\mathcal{A}3+$ is easy to check. Conditions $\mathcal{A}2+$ and $\mathcal{A}7$ are more difficult to check for a general function θ.

Asymptotic Normality

To prove asymptotic normality of our proposed maximum likelihood estimator, we assume $\sigma(t, r(t)) = \sigma(t)$ to facilitate the proofs. The main convenience of this

assumption is the fact that we can find exact distributions of several stochastic integrals with respect to Brownian motion. Before stating our main results, we present the following lemmas that are used in the proof for our theorems.

Lemma 5.4. *Assume that* $|f_T(t)| \leq M_1$ *for* $0 \leq t \leq T, T > 0$. *Then*

$$E\left(\left(\frac{1}{\sqrt{T}}\int_0^T f_T(t)\Big(\int_0^t e^{-a(t-s)}\sigma(s)dW(s)\Big)dt\right)^2\right) \leq \frac{2M_1^2\bar{M}^2}{a^2}.$$

The notation $f_T(t)$ *indicates possible dependence of the function on* T.

One immediate implication of the above lemma is that, for any $\alpha > 0$,

$$\lim_{T\to\infty}\frac{1}{T^{\frac{1}{2}+\alpha}}\int_0^T f_T(t)\Big(\int_0^t e^{-a(t-u)}\sigma(u)dW(u)\Big)dt = 0, \text{ in probability.}$$

To prove this result, we can apply Chebyshev's inequality. We omit the details here.

Lemma 5.5. *Suppose that* $r(t)$ *is the solution to (5.1). We also assume that* $|f_T(t)| \leq C$, *for* $0 \leq t \leq T, T > 0$. *Under the following condition*

($\mathcal{A}8$*)* $\sup_{t>0}\Big|ae^{-at}\int_0^t e^{as}\theta(s)ds - \theta(t)\Big|\sqrt{t} < M$ *for a constant* M,

we have

$$E\left(\left(\frac{1}{\sqrt{T}}\int_0^T f_T(t)\Big(r(t)-\theta(t)\Big)dt\right)^2\right) \leq C^2\Big(3M_a^2r_0^2 + 12M^2 + \frac{6\bar{M}^2}{a^2}\Big), \quad (5.17)$$

where M_a *is is allowed to be dependent on* a.

Note that the right hand-side of (5.17) does not depend on T, i.e., the result is uniform for $T > 0$. Thus, one implication of this lemma is that $\frac{1}{\sqrt{T}}\int_0^T f_T(t)\Big(r(t) - \theta(t)\Big)dt$ is uniformly bounded in probability for $T > 0$. By (5.2), condition $\mathcal{A}8$ essentially requires that the mean of $r(t)$ is close to the mean-reversion level function

$\theta(t)$ in the limit. It is satisfied by any function that converges to its limit faster than $T^{-1/2}$ when $T \to \infty$. This is a technical condition in the sense that it does not affect the set of functions $\theta(t)$ as determined by other assumptions. On any finite interval, $\theta(t)$ may have any shape, and this will not violate the assumption.

The following two theorems are the main results that describe the asymptotic distribution of $\hat{M}_{T,K_T}(\theta)(t)$.

Theorem 5.5. *Assume that condition $\mathcal{A}8$ holds and that the following condition is true:*

($\mathcal{A}9$)

$$\lim_{T\to\infty} \hat{a}_{T,K_T} = a \text{ in probability,}$$

Then we have that, for each $i = 0, 1, \cdots,$

$$\sqrt{T}\hat{a}_{T,K_T}(\hat{\theta}_{i,T} - \theta_{i,T}) \xrightarrow{D} N(0, 2). \tag{5.18}$$

Moreover, for each subset $F_J \in \mathbb{N}$ with finite cardinality J,

$$\sum_{i\in F_J} \frac{T}{2}\hat{a}^2_{T,K_T}(\hat{\theta}_{i,T} - \theta_{i,T})^2 \xrightarrow{D} \chi^2(J). \tag{5.19}$$

The result of this theorem can be used to test the hypothesis that $\theta_{i,T} = 0$, for $i \in J$. We show how to determine the dimension of parameter space in Section 5.4.1.

Theorem 5.6. *Assume that condition $\mathcal{A}8$ holds and that the following condition is true:*

($\mathcal{A}10$)

$$\lim_{T\to\infty} (\hat{a}_{T,K_T} - a)\sqrt{\sum_{i=0}^{K_T} C_i^2} = 0 \text{ in probability,}$$

where C_is are such that $|q_{i,T}(t)| \leq C_i$ uniformly for $0 \leq t \leq T$. For example, we can take $C_i = \sqrt{\frac{2i+1}{2}}$ for the Legendre polynomial system and $C_i = 1$ for the trigonometric polynomial system.

Then we have

$$\sqrt{\frac{T}{\sum_{i=0}^{K_T} q_{i,T}^2(t)}}\hat{a}_{T,K_T}\Big(\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)\Big) \xrightarrow{D} N(0, 2\sigma^2(t)). \tag{5.20}$$

From Theorem 5.6, an approximate 95% confidence interval of $M_{T,K_T}(\theta)(t)$ can be defined as

$$\left(\hat{M}_{T,K_T}(\theta)(t) - \frac{z_{0.975}\sigma(t)\sqrt{\sum_{i=0}^{K_T} 2q_{i,T}^2(t)/T}}{\hat{a}_{T,K}}, \right.$$
$$\left. \hat{M}_{T,K_T}(\theta)(t) - \frac{z_{0.025}\sigma(t)\sqrt{\sum_{i=0}^{K_T} 2q_{i,T}^2(t)/T}}{\hat{a}_{T,K}} \right).$$

Since $\vec{p}_{T,K_T} = \sigma(t)\vec{q}_{T,K_T}$, the above confidence interval is equivalent to the following one:

$$\left(\hat{M}_{T,K_T}(\theta)(t) - \frac{z_{0.975}\sqrt{\sum_{i=0}^{K_T} 2p_{i,T}^2(t)/T}}{\hat{a}_{T,K}}, \right.$$
$$\left. \hat{M}_{T,K_T}(\theta)(t) - \frac{z_{0.025}\sqrt{\sum_{i=0}^{K_T} 2p_{i,T}^2(t)/T}}{\hat{a}_{T,K}} \right). \tag{5.21}$$

We have comments similar to those made at the end of Section 4.3.5 except that now the drift-to-diffusion ratio should be changed from $\frac{\theta(t)}{\sigma(t)}$ to $\frac{a\theta(t)}{\sigma(t)}$.

5.4 Dimension of the Parameter Space and Confidence Intervals for $M_{T,K_T}(\theta)(t)$

In practice, we clearly do not know the true time-dependent level function and hence we risk under-fitting or over-fitting the data by incorporating too few or too many basis functions. In this section, we propose two methods to determine the dimension of parameter space and a confidence interval for $M_{T,K}(\theta)(t)$ for each $t \in [0,T]$. The first one is for large samples, i.e., when T is large. Based on the asymptotic results derived in Section 5.3, we also develop a hypothesis-testing procedure. The idea is the same as the one introduced in Section 4.3.4, except that now we use an asymptotic distribution of the test statistic instead of its exact distribution under the null hypothesis.

The second method that we propose is for small samples. We first find estimates of the mean-reverting speed parameter for K from 0 to K_{max}, and then choose K_{cutoff} such that the mean-reverting speed estimate increases significantly when a parameter space with a dimension higher than K_{cutoff} is employed. This method has been developed based on our simulation studies. It corresponds to the intuition that the mean-reverting speed estimate becomes larger when a higher dimensional parameter space is chosen.

We emphasize that the first method is for large samples and relies on the asymptotic results in Section 5.3. Therefore, it applies to Case 1 only, and $\sigma(t)$ is a deterministic function, i.e., $\vec{p}_{T,K}(t) = \sigma(t)\vec{q}_{T,K}(t)$. The second method can be applied to any sample size, although the method is not justified with full mathematical rigor. Another advantage of the second method is that it applies to both cases, i.e., without constraints on the choice of basis functions or functional form of the volatility function $\sigma(t, r(t))$.

5.4.1 Large Samples

We use an approach similar to the one we developed in Section 4.3.4. Suppose we consider spaces only with dimensions up to K_{max}. For a given number K_0, we are interested in testing the following null hypothesis:

H_0 : *the coefficients corresponding to basis functions of degrees higher than* K_0 *but smaller than* K_{max} *are zero .*

In other words, under H_0, the degree of the time-dependent drift functions is either larger than K_{max} or no larger than K_0. If the null hypothesis is true, then we prefer the more parsimonious parameter space V_{T,K_0} rather than the larger parameter space $V_{T,K_{max}}$.

It follows from Theorem 5.5 that

$$\sum_{i=K_0+1}^{K_T} \frac{T}{2} \hat{a}_{T,K_T}^2 (\hat{\theta}_{i,T}^2 - \theta_{i,T}^2) \xrightarrow{D} \chi^2(K_T - K_0),$$

as $T \to \infty$. However, to use this result, K_T should not be too large for any fixed T. In fact, the results of Theorem 5.5 hold under conditions $\mathcal{A}8$ and $\mathcal{A}9$. Condition $\mathcal{A}8$ is a technical condition. Condition $\mathcal{A}9$ requires that $\hat{a}_{T,K_T}$ converges to a in probability. By Theorem 5.3, condition $\mathcal{A}9$ holds under several assumptions, which require that K_T does not increase too fast with T.

Therefore, if T is large and K_{max} is not too large for a given T, Theorem 5.5 implies that $\sum_{i=K_0+1}^{K_{max}} \frac{T}{2} \hat{a}_{T,K_{max}}^2 (\hat{\theta}_{i,T}^2 - \theta_{i,T}^2)$ can be approximated by $\chi^2(K_{max} - K_0)$. Under H_0, we have that $\theta_{i,T} = 0$, for $i = K_0 + 1, \cdots, K_{max}$. Therefore, $\sum_{i=K_0+1}^{K_{max}} \frac{T}{2} \hat{a}_{T,K_{max}}^2 \hat{\theta}_{i,T}^2$ can be approximated by $\chi^2(K_{max} - K_0)$. Our simulation studies suggest that the hypothesis testing results are reasonable for $T > 50$ and $K_{max} < \sqrt{T}$. We illustrate our hypothesis testing results for large T in Section 5.5.2.

Once the dimension parameter K_T is chosen, we can proceed to finding the confidence interval for $M_{T,K_T}(\theta)(t)$ for each $t \in [0,T]$. For this we use Theorem 5.6. Let $z_{0.975}$ and $z_{0.025}$ be quantiles of a standard normal random variable. Under the conditions in Theorem 5.6, we have

$$\sqrt{\frac{T}{\sum_{i=0}^{K_T} q_{i,T}^2(t)}}\ \hat{a}_{T,K_T}\Big(\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)\Big) \xrightarrow{D} N\Big(0, 2\sigma^2(t)\Big).$$

From (5.21), an approximate 95% confidence interval for $M_{T,K_T}(\theta)(t)$ is

$$\left(\hat{M}_{T,K_T}(\theta)(t) + z_{0.025}\frac{1}{\hat{a}_{T,K_T}}\sqrt{\frac{\sum_{i=0}^{K_T} 2p_{i,T}^2(t)}{T}},\right.$$
$$\left.\hat{M}_{T,K_T}(\theta)(t) + z_{0.975}\frac{1}{\hat{a}_{T,K_T}}\sqrt{\frac{\sum_{i=0}^{K_T} 2p_{i,T}^2(t)}{T}}\right).$$

It can be seen that the width of the confidence interval above is proportional to $\frac{1}{\hat{a}_{T,K_T}}$; the larger $\hat{a}_{T,K}$ is, the narrower the confidence interval for $M_{T,K_T}(\theta)(t)$ is.

5.4.2 Small Samples

The procedure developed in the previous section applies to large sample data sets, i.e., when T is long enough. In practice, available data may not be long enough for this approach to work well. In such cases, we propose using a parametric bootstrap method to correct the bias of $\hat{a}_{T,K_T}$. The bootstrap method has been shown to be an effective method for bias correction and variance estimation for both independent and dependent observations (Efron, 1979; Hall, 1992; Davison and Hinkley, 1997).

We apply the parametric bootstrap method proposed in Tang and Chen (2009) to correct the bias of the mean-reverting speed estimator. Then we determine the dimension of the parameter space by using the following approach:

Step 1: Start from $K = 0$. First we find $\hat{a}_{T,K}$ and $\hat{M}_{T,K}(\theta)(t)$, the maximum likelihood estimators defined by (5.6) and (5.5). Then we apply a basic parametric bootstrap method to correct the bias of $\hat{a}_{T,K}$ and obtain $\hat{a}^{btsrp}_{T,K}$. The bootstrap procedure can be described as follows:

Step (i): Generate a bootstrap sample path $\{r^*(t)\}$ of the same length and with the same sampling interval from the following SDE:

$$dr^*(t) = \hat{a}_{T,K}(\hat{M}_{T,K}(\theta)(t) - r^*(t))dt + \sigma(t, r(t))dW^*(t),$$

where $\sigma(t, r(t))$ is the same as the diffusion coefficient for $\{r(t)\}$.

Step (ii): Obtain new estimators $\hat{a}^*_{T,K}$ and $\hat{M}^*_{T,K}(\theta)(t)$ from the bootstrap sample path by applying the same estimation procedure as for $\hat{a}_{T,K}$ and $\hat{M}_{T,K}(\theta)(t)$.

Step (iii): Repeat Steps i and ii N_B number of times and obtain a set of bootstrap parameter estimates $\hat{a}^{*,b}_{T,K}$ and $\hat{M}^{*,b}_{T,K}(\theta)(t)$ for $b = 1, 2, \cdots, N_B$.

Step (iv): Let $\bar{\hat{a}}^*_{T,K} = N_B^{-1} \sum_{b=1}^{N_B} \hat{a}^{*,b}_{T,K}$. The bootstrap-corrected estimator of "a" is

$$\hat{a}^{btstrp}_{T,K} = 2\hat{a}_{T,K} - \bar{\hat{a}}^*_{T,K}.$$

Step 2: Increase the dimension parameter K and find $\hat{a}_{T,K}$ and $\hat{a}^{btsrp}_{T,K}$.

Step 3: List the estimates $\hat{a}^{btstrp}_{T,K}$ as a function of K. Suppose we can observe a pattern where the estimated values $\hat{a}^{btstrp}_{T,K}$ stabilize near a certain value K_{cutoff} and then increase significantly for $K > K_{cutoff}$. Such a behavior may occur when the data is over fitted, as for large values of K the observed path will tend to revert quickly to the estimated level, leading to high estimates of the reversion speed. If we do not observe such a pattern, then we suggest to take $K_{cutoff} = \lfloor \sqrt{T} \rfloor$, the maximum integer not greater than $\sqrt{T}$. The idea is motivated by conditions $\mathcal{C}1$ and $\mathcal{C}2$ in Theorem 5.4, which are required

for consistency of $\hat{M}_{T,K_T}(\theta)(t)$. In most of our simulation studies, we have observed the above pattern for $\hat{a}_{T,K}^{btstrp}$.

Step 4: Define

$$\begin{aligned}\hat{a}_L &\triangleq \min\{\hat{a}_{T,K}^{btsrp} | \hat{a}_{T,K} > 0, \hat{a}_{T,K}^{btsrp} > 0, K = 0, 1, \cdots, K_{cutoff}\} \\ \hat{a}_U &\triangleq \max\{\hat{a}_{T,K}^{btsrp} | \hat{a}_{T,K} > 0, \hat{a}_{T,K}^{btsrp} > 0, K = 0, 1, \cdots, K_{cutoff}\}.\end{aligned}$$

Then $[\hat{a}_L, \hat{a}_U]$ is the suggested range that contains the true mean-reverting speed parameter a.

With the selected K and $[\hat{a}_L, \hat{a}_U]$, we now proceed to the construction of a confidence interval for $M_{T,K}(\theta)(t)$. From (5.10), a 95% confidence interval for $M_{T,K_T}(\theta)(t)$ is

$$\left(M_{T,K_T}(r)(t) + \frac{\hat{a}_{T,K_T}(\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(r)(t)) - z_{0.975}\sqrt{\vec{p}'_{T,K}(t)A_{T,K}^{-1}\vec{p}_{T,K}(t)}}{a}, \right.$$

$$\left. M_{T,K_T}(r)(t) + \frac{\hat{a}_{T,K_T}(\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(r)(t)) - z_{0.025}\sqrt{\vec{p}'_{T,K}(t)A_{T,K}^{-1}\vec{p}_{T,K}(t)}}{a} \right).$$

Since the confidence interval derived above depends on the unknown parameter "a", we replace a by either $\hat{a}_L$ or $\hat{a}_U$ so that the resulting confidence interval is very likely wider than the true one. For example, if

$$\hat{a}_{T,K_T}\Big(\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(r)(t)\Big) - z_{0.975}\sigma(t, r(t))\sqrt{\vec{q}'_{T,K}(t)A_{T,K}^{-1}\vec{q}_{T,K}(t)} > 0,$$

the left-ending point of the confidence interval for $M_{T,K_T}(\theta)(t)$ is then

$$\frac{\hat{a}_{T,K_T}\Big(\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(r)(t)\Big) - z_{0.975}\sigma(t, r(t))\sqrt{\vec{q}'_{T,K}(t)A_{T,K}^{-1}\vec{q}_{T,K}(t)}}{\hat{a}_U}.$$

Similar argument applies to the right-end point of the confidence interval. For each end point, it is possible to replace a by $\hat{a}_U$ for some t and by $\hat{a}_L$ for other t. Therefore, the resulting confidence bands for $M_{T,K_T}(\theta)(t),\ 0 \le t \le T$ may not be smooth. Simulation examples are presented in Section 5.5.2.

5.5 Simulation Study

In this section, we illustrate our proposed methodology with the following SDE:

$$dr(t) = a\Big(\theta(t) - r(t)\Big)dt + \sigma dW(t), \quad 0 \leq t \leq T, \tag{5.22}$$

which is an example of the general mean-reverting diffusion model (5.1). Because the diffusion coefficient σ is constant, the analytical results in this chapter hold for both cases in Section 5.2. Here we choose the setting in Case 2: $\vec{p}_{T,K} = \vec{q}_{T,K}$. The chosen basis system is Legendre polynomials. For the simulation study through the rest of this chapter we take $a = 0.215$ and $\sigma = 0.0224$. We are interested in estimating the parameters a and $M_{T,K}(\theta)$. In all sections, we assume that $\theta(t) = \frac{0.0672\ L(2,T)}{a}$, where $L(2,T)$ is a Legendre polynomial of degree 2 on $[0,T]$. Hence the drift-to-diffusion ratio is approximately 3 to 1.

The sampling frequency for our simulation study is monthly. We have also implemented simulations with higher frequency, e.g., weekly and daily. However, we have noticed no significant gains for shorter sampling intervals. This finding confirms existing results in the literature, but in different model settings. As shown in Tang and Chen (2009), the estimation accuracy of the parameters in the drift of a stochastic differential equation is mainly driven by the trajectory length rather than the sampling interval. Other similar findings have been reported by CKLS (1992), Stanton (1997), Aït-Sahalia (1999, 2002) and Fan et al. (2003). According to Stanton (1997), as long as data are sampled monthly or more frequently, the errors introduced by using approximations rather than the true drift and diffusion are extremely small when compared with the likely size of estimation errors.

This section is organized as follows:

- In Section 5.5.1, we study the case when the mean-reverting speed parameter a is known. In this case, the asymptotic results in section 5.4.1 are exact.

The estimation results for $M_{T,K(\theta)}$ are as accurate as the ones in Chapter 4, which confirms our analytical results in Section 5.4.1.

- In Section 5.5.2, we estimate $M_{T,K}(\theta)$ when the mean-reverting speed parameter a is unknown. We illustrate the two methods proposed in Section 5.4. One is for large samples and the other is for small samples. For the small-sample case, a basic parametric bootstrap method is proposed to attenuate the bias in estimating the mean-reverting speed parameter. Although we do not have theoretical proof to support applications of bootstrap in our context, our simulation experiences confirm that bootstrap method is effective in correcting the estimation bias for the mean-reverting speed parameter a. This finding is similar to that of Tang and Chen (2009) where a general diffusion process with linear drift is studied. The authors study the case where the mean-reversion level is constant.

5.5.1 When the Mean-Reverting Speed is Known

In the following, we assume that the mean-reverting speed parameter a is known. In this case, it can be verified that the asymptotic results in Theorems 5.5 and 5.6 become exact and condition $\mathcal{A}8$ does not need to hold. Figure 5.1 shows a trajectory of simulated monthly data for $T = 10$ years.

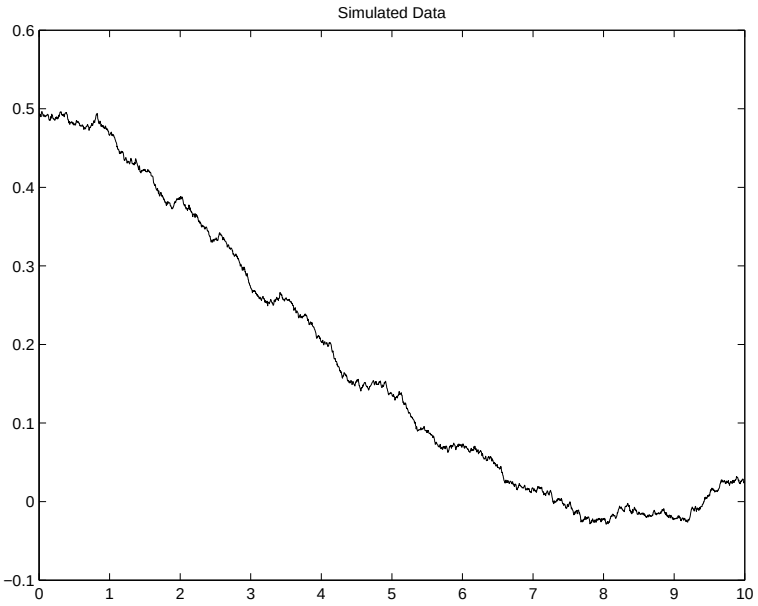


Figure 5.1: Simulated mean-reverting process with a smooth level given by $\theta(t) = 3\sigma \times L(2, 10)/a$, $a = 0.215$, $\sigma = 0.0224$.

We estimate a projection of the level function in the following two steps:

Step 1: To determine the dimension of the parameter space, we use the hypothesis-testing procedure developed in Section 5.4. Table 5.1 shows p-values for our hypothesis testing. The results strongly suggest that the parameter space is spanned by polynomials up to degree 2.

Table 5.1: Hypothesis-testing results when a is known

H_0: the degree of $\theta(t)$ is no larger than	0	1	2	3	4	5	6	7
p-Value (%)	0.0	0.0	85.7	77.2	71.9	64.2	57.2	33.6

Step 2: Based on the above hypothesis-testing result, we set $K = 2$ and

estimate $M_{T=10,K=2}(\theta)$. Figure 5.2 shows the estimation results. One can see that the true level function falls within the 95% confidence interval.

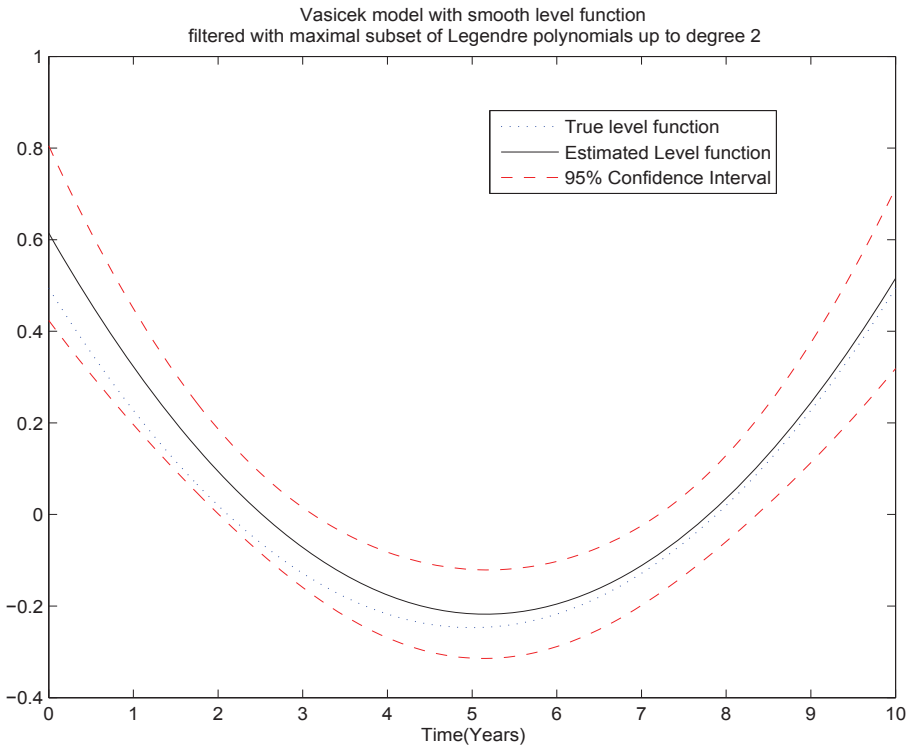


Figure 5.2: Estimation of the level for a Vasicek model with known mean-reverting speed, $a = 0.215$, $\sigma = 0.0224$, $\theta(t) = 3\sigma \times L(2, 10)/a$

5.5.2 When the Mean-Reverting Speed is Unknown

Large Sample

For a large T, we can apply the results developed in Section 5.4.1 to determine the dimension of the parameter space and construct a confidence interval for $M_{T,K_T}(\theta)(t)$. In this section, we set $T = 60$ years.

Step 1: To determine the dimension of the parameter space, we use the hypothesis-testing procedure developed in Section 5.4.1. According to our discussion in Section 5.4.1, we choose $K_{max} = 6 < \sqrt{T}$. Table 5.2 shows the p-values in our testing procedure. The results suggest that the parameter space is spanned by the Legendre polynomials up to degree 2. The choice of K_{max} is still an open question, because the conditions for the asymptotic results in Theorem 5.5 do not provide an explicit connection between K_T and T. We have also implemented other $K_{max} = 3, 4, 5, 6$. The results are consistent. However, when K_{max} increases over 6, the hypothesis-testing procedure tends to choose high-dimension parameter spaces.

Table 5.2: Hypothesis-testing results for large samples

H_0: the degree of $\theta(t)$ is no larger than	0	1	2	3	4	5
p-value (%)	0	0	71.6	55.1	51.9	25.3

Step 2: Based on the above hypothesis-testing result, we set $K = 2$ and estimate $M_{T=60,K=2}(\theta)$. We use the approximate confidence interval for $M_{T,K}(\theta)(t)$ described in Section 5.4.1. Figure 5.3 presents estimation results for $M_{T,K=2}(\theta)$. The estimated mean-reverting speed parameter is $\hat{a}_{T,K=2} = 0.2378$, which is close to the true value $a = 0.215$.

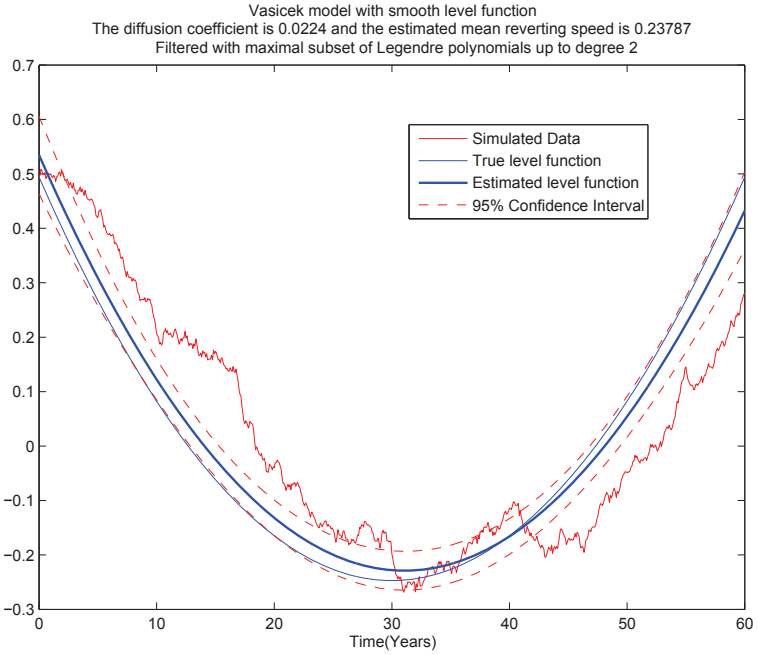


Figure 5.3: Large samples: estimation of the level for a Vasicek model with unknown mean-reverting speed $a = 0.215$, $\sigma = 0.0224$, $\theta(t) = 0.0672 \times L(2, 60)/a$

Small Sample

In this section, we set $T = 10$ and illustrate our methods described in Section 5.4.2. The number of bootstrap re-samples is set equal to 1000, which is the same setting used in Tang and Chen (2009). The following is the procedure for determining the dimension of parameter space and constructing confidence intervals.

Step 1: To determine the dimension of the parameter space, we first find $\hat{a}_{T,K}$ and $\hat{a}^{btstrp}_{T,K}$ for $K = 0, 1, \cdots, 19$. Table 5.3 presents the estimates of the mean-reverting speed a. $\hat{a}^{btstrp}_{T,K}$ is negative for $K = 0, 1$, and it stays in the range $[0.14, 0.59]$ for $K = 3, 4, 5, 6$. Then $\hat{a}^{btstrp}_{T,K}$ increases significantly from 0.14 to 1.13 when K increases from 5 to 6. $\hat{a}^{btstrp}_{T,K}$ is even larger for $K > 7$.

Based on this pattern, we choose $K_{cutoff} = 6$. When $K \leq 6$, the range of a is $[0.14, 0.59]$ based on all cases of $\hat{a}_{T,K}$ and $\hat{a}_{T,K}^{btstrp}$.

Table 5.3: Estimates of mean-reverting speed for different K

K	$\hat{a}_{T,K}$	$\hat{a}_{T,K}^{btstrp}$
0	0.12	-0.42
1	-0.34	-0.40
2	0.55	0.26
3	1.64	0.59
4	1.65	0.30
5	1.96	0.56
6	1.88	0.14
7	2.92	1.13
8	3.71	2.13
9	4.47	2.57
10	5.10	3.05
11	5.40	3.24
12	5.37	2.98
13	5.69	3.27
14	6.55	4.13
15	6.65	4.11
16	7.53	5.00
17	7.96	5.47
18	8.31	5.73
19	8.17	5.54

Step 2: The selected vector space consisting of Legendre polynomials up to degree 6 is employed to estimate a projection of $\theta(t)$. From the results in

Step 1, we set $\hat{a}_L = 0.14$ and $\hat{a}_U = 0.59$. Then we use the approximate 95% confidence interval suggested in Section 5.4.2. Figure 5.4 presents the confidence interval of $M_{T=10,K=6}(\theta)(t)$. The true level function falls within the 95% confidence interval of $M_{T=10,K=6}(\theta)(t)$.

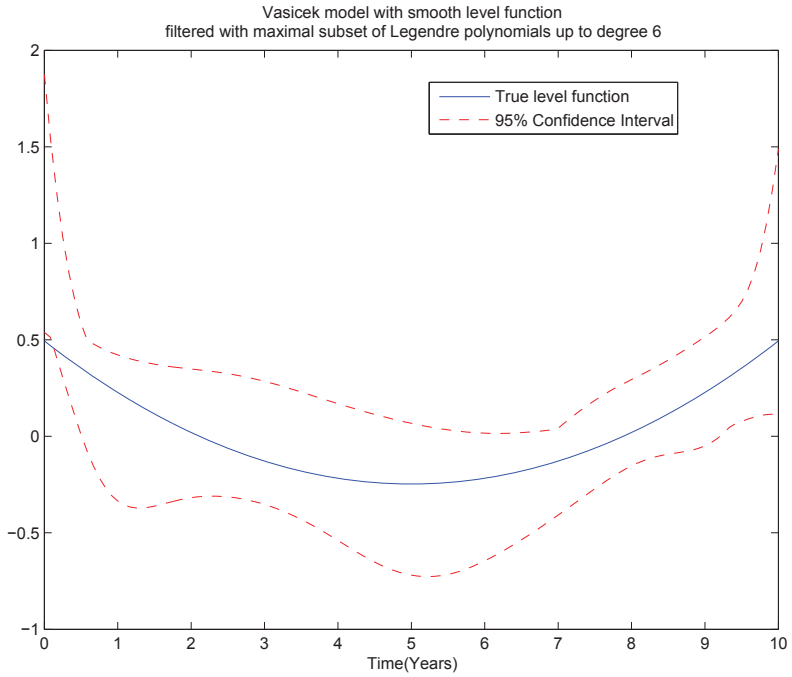


Figure 5.4: Confidence interval for $M_{T,K}(\theta)(t)$: $\sigma = 0.0224$, $a \in$[0.14,0.59], $\theta(t) = 3\sigma \times L(2,10)/a$

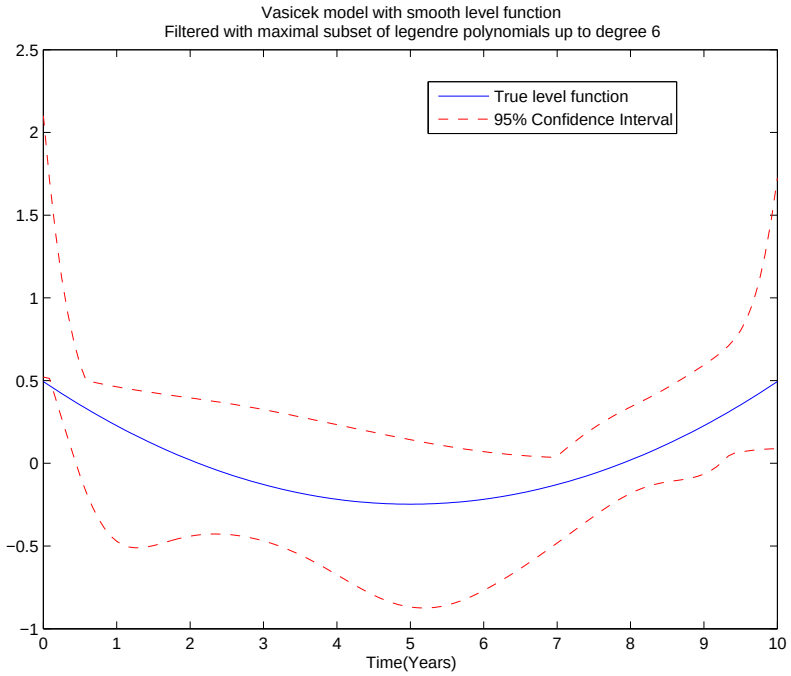


Figure 5.5: Confidence interval for $M_{T,K}(\theta)(t)$: $\sigma = 0.0224$, $a \in$[0.12,1.88], $\theta(t) = 3\sigma \times L(2, 10)/a$

For a different range of a, Figure 5.5 shows the sensitivity of the proposed 95% confidence interval of $M_{T=10,K=6}(\theta)(t)$ to the range of a. In Figure 5.5, $\hat{a}_L = 0.12$ and $\hat{a}_U = 1.88$. Obviously the confidence interval of $M_{T=10,K=6}(\theta)(t)$ becomes larger when the range of a is wider.

5.6 Application to Interest Rates

In this section, we apply model (5.1) and our proposed methodology to the same data set that we investigated in Section 2.5.2. The data set is the same as in CKLS (1992) and consists of monthly observations of annual yields for three-month US T-bills from June 1964 to December 1989. In Section 2.5.2, we fit the data with model (2.5). We detect dramatic movements of the drift in the early 1980s. However, we notice from Figure 5.6 that the estimated volatility[1] based on the quadratic variation of the process is not constant. In particular, the volatility is conspicuous in the early 1980s. It is natural to ask the following question: is the detected dramatic movement in the drift component in the early 1980s a consequence of mis-specification of the volatility (diffusion) function? With the methodology developed in this chapter, we are able to answer this question.

At the end of Section 5.2, we discussed two cases when it comes to the choice of basis functions $\vec{p}_{T,K}(t)$. Both cases result in the same form of 95% confidence interval for $\hat{M}_{T,K}(t)$, as given in (5.10). In particular, Case 2 accommodates the situation when $\sigma(t, r(t)) = \sigma(t)$, i.e. $\sigma(t)$ need not be a constant. Note that the confidence interval in (5.10) depends on the unknown mean-reverting speed parameter 'a'. A range for 'a' can give an approximate 95% confidence interval for $\hat{M}_{T,K}(t)$. In the following, we apply the proposed methodology to determine the number of basis functions K and a range for the parameter 'a'. The bootstrap procedure is the same as described in Section 5.4.2 except that the volatility function $\sigma(t)$ is now estimated from quadratic variation of the real data set rather than being a constant.

More specifically, we first estimate the volatility function from the quadratic variation of the process, which we then use to calculate the matrix $A_{T,K}$ in Section

[1]The volatility function is estimated using a simple formula: $V_t = \sqrt{\frac{(X_{t+\Delta}-X_t)^2}{\Delta}}$, where $\Delta = 1/12$ as we use monthly data.

5.4.2. Then we estimate a projection of the time-dependent level function using the maximum likelihood estimator. The following are the steps taken to estimate a projection of the level function $\theta(\cdot)$:

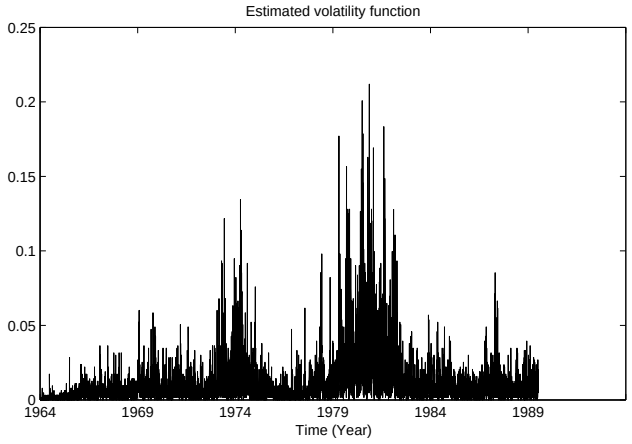


Figure 5.6: Estimated volatility for CKLS data

Step 1: The method for small samples developed in Section 5.4.2 is employed to determine the number of basis functions for estimation. We consider Legendre polynomials of degree from 0 to $K_{max} = 20$. The number of bootstrap resamples is set equal to 1000. Table 5.4 presents the estimation results for a. Note that the bootstrap-corrected mean-reverting speed estimates are mostly negative, which conflicts with the mean-reverting constraint for model (5.1): $a > 0$. Therefore, we focus on $\hat{a}_{T,K}$ in the left column. Based on the testing result, we choose $m = 11$ for estimation of a projection of the level function in Step 2. The range for the estimated mean-reverting speed is $[0.1131, 0.1166]$.

Table 5.4: Interest rate data: estimates of mean-reverting speed for different K

K	$\hat{a}_{T,K}$	$\hat{a}_{T,K}^{btstrp}$
0	0.0054	-0.0210
1	0.0205	-0.0445
2	0.0211	-0.0696
3	0.0437	-0.0716
4	0.0443	-0.1198
5	0.0735	-0.1074
6	0.0254	-0.2064
7	0.0654	-0.1811
8	0.1131	-0.1620
9	0.1149	-0.2182
10	0.1166	-0.2468
11	0.1161	-0.2822
12	0.2004	-0.2072
13	0.1004	-0.3613
14	0.4537	0.0335
15	0.4327	-0.0378
16	0.5051	-0.0061
17	0.5015	-0.0141
18	0.5073	-0.0894
19	0.4931	-0.1413

Step 2: The selected vector space consisting of Legendre polynomials up to degree 11 is employed to estimate the projection of $\theta(t)$. The estimated mean-reverting speed is $\hat{a}_{T,K=11} = 0.1161$. We use the approximate 95% confidence interval suggested in Section 5.4.2.

Figure 5.7 presents the estimation results of $M_{T,K=11}(\theta)$. The results suggest that, after the volatility function is accounted for, there is some evidence of time variation in the mean-reversion level. But at the same time, we cannot exclude the case when our projection is a constant function, because a straight line can be fitted into the 95% confidence interval of the projection. This however may be due to our conservative selection of "a" as explained in Section 5.4.2. There is also some deviation of the data from our model, as the bootstrap results are not consistent with the model. To demonstrate the sensitivity of our result to the estimation of "a", we present Figure 5.8 with confidence interval of the projected mean-reversion level function corresponding to $a = 0.2004$. It can be seen that the confidence interval is narrower compared with Figure 5.7.

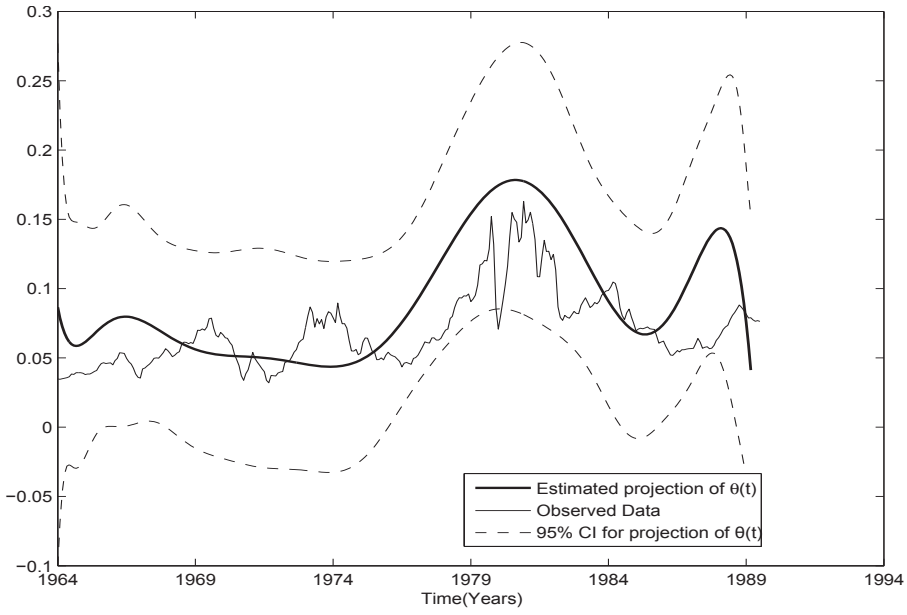


Figure 5.7: Estimation of a projection of the level for interest rate onto a subspace spanned by Legendre polynomials up to degree $K = 11$. $a \in [0.1131, 0.1166]$.

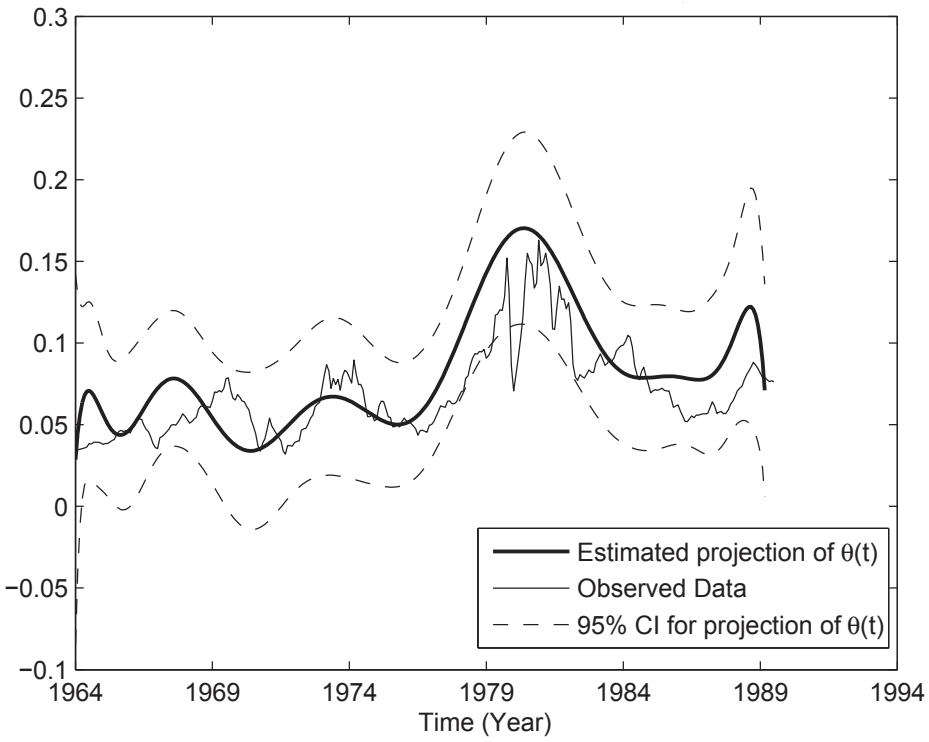


Figure 5.8: Estimation of a projection of the level for interest rate onto a subspace spanned by Legendre polynomials up to degree $K = 12$, $a = 0.2004$.

5.7 Concluding Remarks

In this chapter, we have extended the inference methodology developed in Chapter 4 to a class of mean-reverting time-dependent SDEs. The model (5.1) is more difficult to estimate than model (4.1), since both the mean-reverting speed parameter and the time-dependent level function are unknown. The aliasing problem for estimating both a and $M_{T,K_T}(\theta)(t)$ requires more effort in deriving closed-form maximum likelihood estimator and, in particular, in the proof of asymptotic consistency and normality of the maximum likelihood estimator.

We have proposed two methods for determining the dimension parameter K and finding a confidence interval for the projection $M_{T,K}(\theta)(t)$. When data length T is large, say, at least 50 years of data, we can apply the asymptotic distribution results. When T is moderate, we propose a heuristic approach that combines a basic parametric bootstrap method for selecting K and an approximate 95% confidence interval. Although we have no proof for the second method, we have demonstrated using simulated data that it is easy to implement and works well.

An application of the proposed methodology to the CKLS (1992) data set shows that there is still significant time-variation in the mean-reversion level function when the volatility function is estimated from the quadratic variation of the interest rate process.

In summary, we have demonstrated that the proposed methodology, which is similar in spirit to the sieve method, is both theoretically justifiable and practically implementable. We hope that more research results will be developed for the statistical inference for time-dependent SDEs.

5.8 Appendix: Technical Proofs

Proof of Theorem 5.1

Proof. To solve the system of equations (5.4) and (5.5), we study the numerator and denominator in (5.4) separately. The following expression can be derived from

(5.5) and is useful for studying $\hat{a}_{T,K}$:

$$\begin{aligned}
& \vec{p'}_{T,K}(t)\hat{\Phi}_{T,K}(\theta) - r(t) \\
= \; & \vec{p'}_{T,K}(t)A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)r(t)}{\sigma(t,r(t))^2}dt + \frac{1}{\hat{a}_{T,K}}\vec{p'}_{T,K}(t)A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t) - r(t) \\
= \; & M_{T,K}(r)(t) + \frac{1}{\hat{a}_{T,K}}\vec{p'}_{T,K}(t)A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t) - r(t) \\
= \; & \frac{1}{\hat{a}_{T,K}}\vec{p'}_{T,K}(t)A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t) - M_{T,K}^{\perp}(r)(t), \qquad (5.23)
\end{aligned}$$

where the second equation follows from the definition of $M_{T,K}$, and the last equation follows from the fact that $M_{T,K}^{\perp}(r)(t) = r(t) - M_{T,K}(r)(t)$.

By (5.23), the numerator of (5.4) can be rewritten as follows:

$$\begin{aligned}
& \int_0^T \frac{[\vec{p'}_{T,K}(t)\hat{\Phi}_{T,K}(\theta) - r(t)]}{\sigma(t,r(t))^2}dr(t) \\
= \; & \frac{1}{\hat{a}_{T,K}}\Big(\int_0^T \frac{\vec{p'}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t)\Big)A_{T,K}^{-1}\Big(\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t)\Big) \\
& - \int_0^T \frac{M_{T,K}^{\perp}(r)(t)}{\sigma(t,r(t))^2}dr(t). \qquad (5.24)
\end{aligned}$$

Using (5.23) again and noting that $\{p_{i,T}, 0 \le i \le K\}$ are orthogonal to $M_{T,K}^{\perp}(r)(t)$

in $L^2([0,T],\mu(dt))$, we can rewrite the denominator of (5.4):

$$\begin{aligned}
&\int_0^T \frac{[\vec{p'}_{T,K}(t)\hat{\Phi}_{T,K}(\theta)-r(t)]^2}{\sigma(t,r(t))^2}dt \\
= \;& \frac{1}{\hat{a}_{T,K}^2}\int_0^T \frac{[\vec{p'}_{T,K}(t)A_{T,K}^{-1}\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t)]^2}{\sigma(t,r(t))^2}dt+\int_0^T \frac{[M_{T,K}^{\perp}(r)(t)]^2}{\sigma(t,r(t))^2}dt \\
= \;& \frac{1}{\hat{a}_{T,K}^2}\int_0^T \frac{(\int_0^T \frac{\vec{p'}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t))A_{T,K}^{-1}\vec{p}_{T,K}(t)\vec{p'}_{T,K}(t)A_{T,K}^{-1}(\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t))}{\sigma(t,r(t))^2}dt \\
&+\int_0^T \frac{[M_{T,K}^{\perp}(r)(t)]^2}{\sigma(t,r(t))^2}dt \\
= \;& \frac{1}{\hat{a}_{T,K}^2}(\int_0^T \frac{\vec{p'}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t))A_{T,K}^{-1}(\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t)) \\
&+\int_0^T \frac{[M_{T,K}^{\perp}(r)(t)]^2}{\sigma(t,r(t))^2}dt. \qquad (5.25)
\end{aligned}$$

Therefore, by combining (5.24) and (5.25), (5.4) becomes

$$\hat{a}_{T,K}=\frac{\frac{1}{\hat{a}_{T,K}}(\int_0^T \frac{\vec{p'}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t))A_{T,K}^{-1}(\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t))-\int_0^T \frac{M_{T,K}^{\perp}(r)(t)}{\sigma(t,r(t))^2}dr(t)}{\frac{1}{\hat{a}_{T,K}^2}(\int_0^T \frac{\vec{p'}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t))A_{T,K}^{-1}(\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t))+\int_0^T \frac{[M_{T,K}^{\perp}(r)(t)]^2}{\sigma(t,r(t))^2}dt}.$$

Multiplying both sides of the above equation by the denominator on the right side, we have

$$\begin{aligned}
&\frac{1}{\hat{a}_{T,K}}(\int_0^T \frac{\vec{p'}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t))A_{T,K}^{-1}(\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t))+\hat{a}_{T,K}\int_0^T \frac{[M_{T,K}^{\perp}(r)(t)]^2}{\sigma(t,r(t))^2}dt \\
= \;& \frac{1}{\hat{a}_{T,K}}(\int_0^T \frac{\vec{p'}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t))A_{T,K}^{-1}(\int_0^T \frac{\vec{p}_{T,K}(t)}{\sigma(t,r(t))^2}dr(t))-\int_0^T \frac{M_{T,K}^{\perp}(r)(t)}{\sigma(t,r(t))^2}dr(t).
\end{aligned}$$

Canceling the first term on both sides of the above equation leads to

$$\hat{a}_{T,K}=\frac{-\int_0^T \frac{M_{T,K}^{\perp}(r)(t)}{\sigma(t,r(t))^2}dr(t)}{\int_0^T \frac{[M_{T,K}^{\perp}(r)(t)]^2}{\sigma(t,r(t))^2}dt}.$$

□

Proof of Theorem 5.3

From (5.6) and (5.1),

$$\begin{aligned}
& \hat{a}_{T,K_T} \\
= & \frac{-\int_0^T \frac{M^{\perp}_{T,K_T}(r)(t)}{\sigma(t,r(t))^2} dr(t)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2 \mu(dt)} \\
= & \frac{-\int_0^T \frac{M^{\perp}_{T,K_T}(r)(t)}{\sigma(t,r(t))^2} [a(\theta(t) - r(t))dt + \sigma(t,r(t))dW(t)]}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2 \mu(dt)} \\
= & a \cdot \frac{\int_0^T M^{\perp}_{T,K_T}(r)(t) r(t) \mu(dt)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2 \mu(dt)} - a \cdot \frac{\int_0^T M^{\perp}_{T,K_T}(r)(t) \theta(t) \mu(dt)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2 \mu(dt)} - \frac{\int_0^T \frac{M^{\perp}_{T,K_T}(r)(t)}{\sigma(t,r(t))} dW(t)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2 \mu(dt)} \\
= & a - a \cdot \frac{\int_0^T M^{\perp}_{T,K_T}(r)(t) M^{\perp}_{T,K_T}(\theta)(t) \mu(dt)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2 \mu(dt)} - \frac{\int_0^T \frac{M^{\perp}_{T,K_T}(r)(t)}{\sigma(t,r(t))} dW(t)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2 \mu(dt)}. \qquad (5.26)
\end{aligned}$$

In the following, we prove that both the second and third terms of (5.26) converge to zero in probability. Then the conclusion of Theorem 5.3 will follow immediately from Slutsky's theorem.

For the second term on the right-hand side of equation (5.26), the Cauchy-Schwartz inequality implies that

$$\left| \frac{\int_0^T M^{\perp}_{T,K_T}(r)(t) M^{\perp}_{T,K_T}(\theta)(t) \mu(dt)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2 \mu(dt)} \right| \leq \sqrt{\frac{\int_0^T [M^{\perp}_{T,K_T}(\theta)(t)]^2 \mu(dt)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2 \mu(dt)}}. \qquad (5.27)$$

This result and condition $\mathcal{A}2$ imply that

$$\lim_{T \to \infty} \left| \frac{\int_0^T M^{\perp}_{T,K_T}(r)(t) M^{\perp}_{T,K_T}(\theta)(t) \mu(dt)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2 \mu(dt)} \right| = 0, \ \textit{in probability.}$$

To show that the third term in (5.26) converges to zero in probability, we first

represent it as follows

$$
\begin{aligned}
&\frac{\int_0^T \frac{M^{\perp}_{T,K_T}(r)(t)}{\sigma(t,r(t))}dW(t)}{\int_0^T [M^{\perp}_{T,K_T}(r)(t)]^2\mu(dt)} \\
= \;& \frac{\int_0^T \frac{r(t)}{\sigma(t,r(t))}dW(t) - \int_0^T \frac{M_{T,K_T}(r)(t)}{\sigma(t,r(t))}dW(t)}{\int_0^T r(t)^2\mu(dt) - \int_0^T [M_{T,K_T}(r)(t)]^2\mu(dt)} \\
= \;& \frac{\int_0^T \frac{r(t)}{\sigma(t,r(t))}dW(t) - \int_0^T \frac{\vec{p'}_{T,K_T}(t)A^{-1}_{T,K}\int_0^T \vec{p}_{T,K_T}(t)r(t)\mu(dt)}{\sigma(t,r(t))}dW(t)}{\int_0^T r(t)^2\mu(dt) - \int_0^T [M_{T,K_T}(r)(t)]^2\mu(dt)} \\
= \;& \frac{\int_0^T \frac{r(t)}{\sigma(t,r(t))}dW(t) - [\int_0^T \vec{p'}_{T,K_T}(t)r(t)\mu(dt)]A^{-1}_{T,K}[\int_0^T \frac{\vec{p}_{T,K_T}(t)}{\sigma(t,r(t))}dW(t)]}{\int_0^T r(t)^2\mu(dt) - \int_0^T [M_{T,K_T}(r)(t)]^2\mu(dt)} \\
= \;& \frac{\frac{\int_0^T \frac{r(t)}{\sigma(t,r(t))}dW(t)}{\int_0^T r(t)^2\mu(dt)} - \frac{[\int_0^T \vec{p'}_{T,K_T}(t)r(t)\mu(dt)]A^{-1}_{T,K}[\int_0^T \frac{\vec{p}_{T,K_T}(t)}{\sigma(t,r(t))}dW(t)]}{\int_0^T r(t)^2\mu(dt)}}{1 - \frac{\int_0^T [M_{T,K_T}(r)(t)]^2\mu(dt)}{\int_0^T r(t)^2\mu(dt)}},
\end{aligned}
$$

where in the third line we used the definition of $M_{T,K_T}(r)(t)$, and in the last line we divided both the numerator and denominator by $\int_0^T r^2(t)\mu(dt)$.

Under condition $\mathcal{A}1$, we have that with probability one the denominator $1 - \frac{\int_0^T [M_{T,K_T}(r)(t)]^2\mu(dt)}{\int_0^T r(t)^2\mu(dt)}$ is bounded away from 0 when T is large enough. Therefore, we need to show only that the numerator

$$
\begin{aligned}
&\frac{\int_0^T \frac{r(t)}{\sigma(t,r(t))}dW(t)}{\int_0^T r(t)^2\mu(dt)} - \frac{[\int_0^T \vec{p'}_{T,K_T}(t)r(t)\mu(dt)]A^{-1}_{T,K}[\int_0^T \frac{\vec{p}_{T,K_T}(t)}{\sigma(t,r(t))}dW(t)]}{\int_0^T r(t)^2\mu(dt)} \\
&\xrightarrow{P} 0, \;\; as \; T \to \infty. \qquad (5.28)
\end{aligned}
$$

We prove (5.28) by showing that both terms converge to 0 in probability and then apply Slutsky's theorem.

For the first term of (5.28), it follows from the definition $\mu(dt) = \frac{1}{\sigma^2(t,r(t))}dt$ and condition $\mathcal{A}3$ that $P(\int_0^\infty \frac{r(t)^2}{\sigma^2(t,r(t))}dt = \infty) = 1$. Then Lemma 5.1 implies that

$$
P(\lim_{T\to\infty} \frac{\int_0^T \frac{r(t)}{\sigma(t,r(t))}dW(t)}{\int_0^T r(t)^2\mu(dt)} = 0) = 1.
$$

Therefore, the first term in (5.28) converges to 0 in probability.

To show that the second term in (5.28), given by

$$\frac{[\int_0^T \vec{p'}_{T,K_T}(t)r(t)\mu(dt)]A_{T,K}^{-1}[\int_0^T \frac{\vec{p}_{T,K_T}(t)}{\sigma(t,r(t))}dW(t)]}{\int_0^T r(t)^2\mu(dt)}, \tag{5.29}$$

converges to 0 in probability, we first study its numerator:

$$\begin{aligned}
& \left|[\int_0^T \vec{p'}_{T,K_T}(t)r(t)\mu(dt)]A_{T,K}^{-1}[\int_0^T \frac{\vec{p}_{T,K_T}(t)}{\sigma(t,r(t))}dW(t)]\right| \\
= & \frac{2}{T}\left|\int_0^T \vec{p'}_{T,K_T}(t)r(t)\mu(dt)\int_0^T \frac{\vec{p}_{T,K_T}(t)}{\sigma(t,r(t))}dW(t)\right| \\
= & \frac{2}{T}\left|\sum_{i=0}^{K_T}\int_0^T p_{i,T}(t)r(t)\mu(dt)\int_0^T \frac{p_{i,T}(t)}{\sigma(t,r(t))}dW(t)\right| \\
= & \frac{2}{T}\left|\sum_{i=0}^{K_T}\int_0^T p_{i,T}(t)r(t)\mu(dt)\int_0^T q_{i,T}(t)dW(t)\right| \\
\leq & \frac{2}{T}\sqrt{\sum_{i=0}^{K_T}(\int_0^T p_{i,T}(t)r(t)\mu(dt))^2\sum_{i=0}^{K_T}[\int_0^T q_{i,T}(t)dW(t)]^2},
\end{aligned} \tag{5.30}$$

where in the second line we used the fact that $A_{T,K} = \frac{T}{2}I_{(K+1)\times(K+1)}$ and in the last line we used the Cauchy-Schwartz inequality. From the definition of the orthogonal projection operator M_{T,K_T}, as in Appendix 4.7.3, we have

$$\begin{aligned}
M_{T,K_T}(r)(t) &= \vec{p'}_{T,K_T}(t)A_{T,K}^{-1}\int_0^T \vec{p}_{T,K_T}(t)r(t)\mu(dt) \\
&= \sum_{i=0}^{K_T}\frac{2}{T}p_{i,T}(t)\int_0^T p_{i,T}(t)r(t)\mu(dt).
\end{aligned}$$

Therefore,

$$\int_0^T [M_{T,K_T}(r)(t)]^2\mu(dt) = \frac{2}{T}\sum_{i=0}^{K_T}[\int_0^T p_{i,T}(t)r(t)\mu(dt)]^2, \tag{5.31}$$

where we used the fact that $\{\vec{p}_{i,T}\}_{i\geq 0}$ is an orthogonal basis for $L^2([0,T],\mu(dt))$.

Thus it follows from (5.30) and (5.31) that

$$\begin{aligned} & \left| [\int_0^T \vec{p'}_{T,K_T}(t) r(t) \mu(dt)] A_{T,K}^{-1} [\int_0^T \frac{\vec{p}_{T,K_T}(t)}{\sigma(t,r(t))} dW(t)] \right| \\ \leq & \sqrt{\int_0^T [M_{T,K_T}(r)(t)]^2 \mu(dt) \cdot \frac{2}{T} \sum_{i=0}^{K_T} [\int_0^T q_{i,T}(t) dW(t)]^2}. \end{aligned} \tag{5.32}$$

Denote $\frac{2}{T} \sum_{i=0}^{K_T} [\int_0^T q_i(t) dW(t)]^2$ by ξ. It follows from the above inequality and (5.29) that

$$\begin{aligned} & \frac{[\int_0^T \vec{p'}_{T,K_T}(t) r(t) \mu(dt)] A_{T,K}^{-1} [\int_0^T \frac{\vec{p}_{T,K_T}(t)}{\sigma(t,r(t))} dW(t)]}{\int_0^T r(t)^2 \mu(dt)} \\ \leq & \frac{\sqrt{\int_0^T [M_{T,K_T}(r)(t)]^2 \mu(dt)} \cdot \sqrt{\xi}}{\int_0^T r(t)^2 \mu(dt)} \\ \leq & \sqrt{\frac{\xi}{\int_0^T r(t)^2 \mu(dt)}}. \end{aligned}$$

We now show that, under condition $(\mathcal{A}3)$, we have $\lim_{T\to\infty} \frac{\xi}{\int_0^T r(t)^2 \mu(dt)} = 0$ in probability. This will prove that (5.29) converges to zero in probability and will conclude our proof. For any $\epsilon, \eta > 0$,

$$\begin{aligned} & P(\frac{\xi}{\int_0^T r(t)^2 \mu(dt)} > \epsilon) \\ \leq & P(\frac{K_T + 1}{\int_0^T r(t)^2 \mu(dt)} > \eta) + P(\frac{\xi}{K_T + 1} > \frac{\epsilon}{\eta}), \\ \leq & P(\frac{K_T + 1}{\int_0^T r(t)^2 \mu(dt)} > \eta) + \frac{\eta}{\epsilon(K_T + 1)} E(\xi), \end{aligned} \tag{5.33}$$

where in (5.33) we used Markov Inequality.

We show that $E(\xi) = K_T + 1$. The definition of $\{q_{i,T}(t)\}$ in Section 4.3.1 implies that $\{\sqrt{\frac{2}{T}} q_{i,T}(t), i \geq 0\}$ is an orthonormal basis for $L^2([0,T], dt)$. Therefore, $\{\int_0^T \sqrt{\frac{2}{T}} q_{i,T}(t) dW(t)\}$ are independent standard normal random variables, which

implies

$$\xi = \frac{2}{T}\sum_{i=0}^{K_T}[\int_0^T q_i(t)dW(t)]^2 \sim \chi^2(K_T+1).$$

Then $E[\xi] = K_T + 1$ and (5.33) imply that for any $\epsilon, \eta > 0$,

$$\begin{aligned} & P(\frac{\xi}{\int_0^T r(t)^2\mu(dt)} > \epsilon) \\ \leq \quad & P(\frac{K_T+1}{\int_0^T r(t)^2\mu(dt)} > \eta) + \frac{\eta}{\epsilon} \\ \underrightarrow{P} \quad & 0 + \frac{\eta}{\epsilon}, \end{aligned}$$

where the last line follows from condition $\mathcal{A}3$. The arbitrariness of η implies that $\lim_{T\to\infty} P(\frac{\xi}{\int_0^T r(t)^2\mu(dt)} > \epsilon) = 0$. □

Proof of Lemma 5.2

By the definition of the operator $M_{T,K}$,

$$|M_{T,K}(f)(t)| = \vec{p}_{T,K}(t)'\Phi_{T,K}(f) = \vec{p}_{T,K}(t)'A_{T,K}^{-1}\int_0^T \vec{p}_{T,K}(t)f(t)\mu(dt).$$

Using the above result and the fact that $A_{T,K}^{-1} = \frac{2}{T}I_{(K_T+1)\times(K_T+1)}$, we have

$$\begin{aligned} & |M_{T,K}(f)(t)| \\ = \quad & \frac{2}{T}|\sum_{j=0}^{K_T} p_{j,T}(t)\int_0^T p_{j,T}(t)f(t)\mu(dt)| \\ \leq \quad & \frac{2}{T}\sqrt{\sum_{j=0}^{K_T} p_{j,T}^2(t)\sum_{j=0}^{K_T}(\int_0^T p_{j,T}(t)f(t)\mu(dt))^2}, \\ = \quad & \frac{2\sigma(t,r(t))}{T}\sqrt{\sum_{j=0}^{K_T} q_{j,T}^2(t)}\sqrt{\sum_{j=0}^{K_T}(\int_0^T p_{j,T}(t)f(t)\mu(dt))^2}, \end{aligned}$$

where in the second last line we used the Cauchy-Schwartz inequality, and the last line holds for $p_{j,T}(t) = \sigma(t, r(t))q_{j,T}(t)$.

Since $p_{j,T}$ are orthogonal vectors in $L^2([0,T],\mu(dt))$, we have

$$\begin{aligned}
& \frac{2\sigma(t,r(t))}{T}\sqrt{\sum_{j=0}^{K_T} q_{j,T}^2(t)}\sqrt{\sum_{j=0}^{K_T}(\int_0^T p_{j,T}(t)f(t)\mu(dt))^2} \\
\leq\ & \frac{2\sigma(t,r(t))}{T}\sqrt{\sum_{j=0}^{K_T} q_{j,T}^2(t)}\sqrt{\sum_{j=0}^{K_T}\frac{T}{2}\int_0^T f^2(t)\mu(dt)}, \\
\leq\ & \frac{2\sigma(t,r(t))}{T}\sqrt{\sum_{j=0}^{K_T} q_{j,T}^2(t)}\sqrt{\sum_{j=0}^{K_T}\frac{T}{2}\int_0^T \frac{f^2(t)}{\bar{\epsilon}^2}dt}, \\
=\ & \frac{\sigma(t,r(t))}{\bar{\epsilon}}\sqrt{\frac{2\int_0^T f^2(t)dt}{T}}\sqrt{\sum_{j=0}^{K_T} q_{j,T}^2(t)},
\end{aligned}$$

where the second last line holds because $\mu(dt) = \frac{1}{\sigma^2(t,r(t))}dt$ and by our assumption $0 < \bar{\epsilon} < \sigma(t,r(t)) < \bar{M}$.

Proof of Lemma 5.3

The idea is to decompose $r(t) - \theta(t)$ into a sum of a deterministic function and a random function. By (5.2),

$$r(t) = e^{-at}r_0 + \int_0^t e^{-a(t-u)}\sigma(u,r(u))dW(u) + a\int_0^t e^{-a(t-u)}\theta(u)du.$$

Then we have

$$\begin{aligned}
& \frac{1}{T}\int_0^T (r(t)-\theta(t))^2 dt \\
=\ & \frac{1}{T}\int_0^T \Big(e^{-at}r_0 + a\int_0^t e^{-a(t-u)}\theta(u)du - \theta(t) + \int_0^t e^{-a(t-u)}\sigma\big(u,r(u)\big)dW(u)\Big)^2 dt \\
\leq\ & \frac{2}{T}\int_0^T \Big(e^{-at}r_0 + a\int_0^t e^{-a(t-u)}\theta(u)du - \theta(t)\Big)^2 dt + \frac{2}{T}\int_0^T \Big(\int_0^t e^{-a(t-u)}\sigma(u,r(u))dW(u)\Big)^2 dt.
\end{aligned}$$

Itô's isometry and the Fubini's theorem imply that

$$\begin{aligned}
& E[\frac{1}{T}\int_0^T (r(t)-\theta(t))^2 dt] \\
\leq\ & \frac{2}{T}\int_0^T (e^{-at}r_0 + a\int_0^t e^{-a(t-u)}\theta(u)du - \theta(t))^2 dt + \frac{2}{T}\int_0^T (\int_0^t e^{-a(t-u)}E[\sigma^2(u,r(u))]du)^2 dt.
\end{aligned}$$

Now we show respectively that the two terms on the right hand-side of the above inequality are uniformly bounded for $T > 0$. For the first term, since $|\theta(t)| \leq \bar{M}$,

$$\begin{aligned} & \frac{2}{T}\int_0^T (e^{-at}r_0 + a\int_0^t e^{-a(t-u)}\theta(u)du - \theta(t))^2 dt \\ \leq \; & \frac{2}{T}\int_0^T (e^{-at}|r_0| + ae^{-at}\bar{M}\frac{e^{at}-1}{a} + \bar{M})dt \leq \frac{2}{T}\int_0^T (2\bar{M} + |r_0|)dt = 4\bar{M} + 2|r_0|, \end{aligned}$$

which is uniformly bounded for $T > 0$. For the second term, since $|\sigma(t, r(t))| \leq \bar{M}$,

$$\begin{aligned} & \frac{2}{T}\int_0^T (\int_0^t e^{-a(t-u)}E[\sigma^2(u, r(u))]du)^2 dt \\ \leq \; & \frac{2}{T}\int_0^T \bar{M}^2 e^{-at}(\int_0^t e^{2au})dt \\ = \; & \frac{2\bar{M}^2}{T}\int_0^T e^{-2at}\frac{e^{2at}-1}{2a}dt \\ = \; & \frac{\bar{M}^2}{aT}\int_0^T (1 - e^{-2at})dt \leq \bar{M}^2, \end{aligned}$$

which is uniformly bounded for $T > 0$. Thus we have proved (5.16). □

Proof of Theorem 5.4

From (5.12), we know that

$$\begin{aligned} & \frac{\hat{a}_{T,K_T}[\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)]}{\sigma(t, r(t))} + \frac{(\hat{a}_{T,K_T} - a)M_{T,K_T}(\theta - r)(t)}{\sigma(t, r(t))} \\ \sim \; & N(0, \vec{q}'_{T,K_T}(t)A^{-1}_{T,K_T}\vec{q}_{T,K_T}(t)) \\ = \; & N(0, \frac{2}{T}\sum_{i=0}^{K_T} q^2_{i,T}(t)). \end{aligned}$$

The same argument as in the proof of Theorem 4.2 implies that under conditions $\mathcal{C}1$ and $\mathcal{C}2$, the term

$$\hat{a}_{T,K_T}[\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)] + (\hat{a}_{T,K_T} - a)M_{T,K_T}(\theta - r)(t)$$

converges to zero in probability. Since it follows from condition $\mathcal{A}6$ that $\lim_{T\to\infty} \hat{a}_{T,K_T} = a$ in probability, the statement of this theorem will follow as long as we prove that $(\hat{a}_{T,K_T} - a)M_{T,K_T}(\theta - r)(t)$ converges to 0 in probability. In light of the following expression,

$$(\hat{a}_{T,K_T} - a)M_{T,K_T}(\theta - r)(t) = (\hat{a}_{T,K_T} - a)\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)}\frac{M_{T,K_T}(\theta - r)(t)}{\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)}}$$

and condition $\mathcal{A}6$, a sufficient condition for this result to hold is that $\frac{M_{T,K_T}(\theta-r)(t)}{\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)}}$ is uniformly bounded in probability for $T > 0$. Indeed, by Lemma 5.2, we have

$$\frac{|M_{T,K_T}(\theta - r)(t)|}{\sqrt{\sum_{j=0}^{K_T} q_{j,T}^2(t)}} \leq \frac{\sigma(t, r(t))}{\epsilon}\sqrt{\frac{2\int_0^T (\theta(t) - r(t))^2 dt}{T}}.$$

On the other hand, from Lemma 5.3, $\frac{2\int_0^T (\theta(t)-r(t))^2 dt}{T}$ is uniformly bounded in probability for $T > 0$. Therefore, $\frac{|M_{T,K_T}(\theta-r)(t)|}{\sqrt{\sum_{j=0}^{K_T} q_{j,T}^2(t)}}$ is also uniformly bounded in probability for $T > 0$. □

Proof of Lemma 5.4

The idea is to write $(\frac{1}{\sqrt{T}}\int_0^T f_T(t)(\int_0^t e^{-a(t-s)}\sigma(s)dW(s))dt)^2$ as a double integral and then take expectation inside the double integral. In the definition, we also use the Itô isometry and Fubini's theorem. The details are as follows:

$$\begin{aligned}
& E[(\frac{1}{\sqrt{T}}\int_0^T f_T(t)(\int_0^t e^{-a(t-u)}\sigma(u)dW(u))dt)^2] \\
= \; & \frac{1}{T}E[\int_0^T\int_0^T f_T(t)f_T(s)e^{-a(t+s)}(\int_0^t e^{au}\sigma(u)dW(u)\int_0^s e^{av}\sigma(v)dW_v)dsdt] \\
= \; & \frac{1}{T}\int_0^T\int_0^T f_T(t)f_T(s)e^{-a(t+s)}E(\int_0^t e^{au}\sigma(u)dW(u)\int_0^s e^{av}\sigma(v)dW_v)dsdt \\
= \; & \frac{1}{T}E[\int_0^T\int_0^T f_T(t)f_T(s)e^{-a(t+s)}(\int_0^{\min(t,s)}\sigma^2(u)e^{2au}du)dsdt]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{M_1^2}{T}\int_0^T\int_0^T e^{-a(t+s)}(\int_0^{\min(t,s)} e^{2au}\sigma^2(u)du)dsdt \\
&= \frac{M_1^2\bar{M}^2}{T}\int_0^T\int_0^t e^{-a(t+s)}\frac{e^{2as}-1}{2a}dsdt + \frac{M_1^2\bar{M}^2}{T}\int_0^T\int_t^T e^{-a(t+s)}\frac{e^{2at}-1}{2a}dsdt \\
&= \frac{M_1^2\bar{M}^2}{2Ta^2}\int_0^T(1+e^{-2at}-2e^{-at})dt + \frac{M_1^2\bar{M}^2}{2Ta^2}\int_0^T(1-e^{-2at}-e^{a(t-T)}+e^{-a(T+t)})dt \\
&\leq \frac{M_1^2\bar{M}^2}{a^2}+\frac{M_1^2\bar{M}^2}{a^2} = \frac{2M_1^2\bar{M}^2}{a^2},
\end{aligned}
$$

where we have omitted some elementary transformations. □

Proof of Lemma 5.5

The idea is to decompose $\frac{1}{\sqrt{T}}\int_0^T f_T(t)(r(t)-\theta(t))dt$ into three terms and consider them separately. The following are the details. By (5.2),

$$
r(t)-\theta(t) = e^{-at}r_0 + a\int_0^t e^{-a(t-u)}\theta(u)du - \theta(t) + \int_0^t e^{-a(t-u)}\sigma(u)dW(u).
$$

Then,

$$
\begin{aligned}
&E[(\frac{1}{\sqrt{T}}\int_0^T f_T(t)(r(t)-\theta(t))dt)^2] \\
\leq\ & 3(\frac{1}{\sqrt{T}}\int_0^T f_T(t)e^{-at}r_0dt)^2 + 3(\frac{1}{\sqrt{T}}\int_0^T f_T(t)(a\int_0^t e^{-a(t-u)}\theta(u)du-\theta(t))dt)^2 + \\
& 3E[(\frac{1}{\sqrt{T}}\int_0^T f_T(t)(\int_0^t e^{-a(t-u)}\sigma(u)dW(u))dt)^2]
\end{aligned}
$$

We consider the three terms on the right hand-side of the above inequality separately. For the first term,

$$
\begin{aligned}
&|\frac{1}{\sqrt{T}}\int_0^T f_T(t)e^{-at}r_0dt| \\
\leq\ & C|r_0|\frac{1}{\sqrt{T}}\frac{1-e^{-aT}}{a} \leq M_aC|r_0|,
\end{aligned}
$$

where $M_a \triangleq \max_{T>0}\frac{1}{\sqrt{T}}\frac{1-e^{-aT}}{a}$. The maximum of $\frac{1}{\sqrt{T}}\frac{1-e^{-aT}}{a}$ exists on $[0,\infty]$ because

$$
\lim_{T\to 0}\frac{1}{\sqrt{T}}\frac{1-e^{-aT}}{a} = \lim_{T\to\infty}\frac{1}{\sqrt{T}}\frac{1-e^{-aT}}{a} = 0.
$$

Therefore,

$$3(\frac{1}{\sqrt{T}}\int_0^T f_T(t)e^{-at}r_0dt)^2 \leq 3M_a^2C^2r_0^2.$$

For the second term, under the condition $\mathcal{A}8$, we have

$$\begin{aligned} & |\frac{1}{\sqrt{T}}\int_0^T f_T(t)(a\int_0^t e^{-a(t-u)}\theta(u)du - \theta(t))dt| \\ \leq & \frac{1}{\sqrt{T}}\int_0^T C\frac{M}{\sqrt{t}}dt = 2MC. \end{aligned}$$

Therefore,

$$3(\frac{1}{\sqrt{T}}\int_0^T f_T(t)(a\int_0^t e^{-a(t-u)}\theta(u)du - \theta(t))dt)^2 \leq 12M^2C^2.$$

For the third term, by Lemma 5.4,

$$3E[(\frac{1}{\sqrt{T}}\int_0^T f_T(t)(\int_0^t e^{-a(t-u)}\sigma(u)dW(u))dt)^2] \leq \frac{6C^2\bar{M}^2}{a^2}.$$

Adding all three terms together, we get

$$\begin{aligned} & E[(\frac{1}{\sqrt{T}}\int_0^T f_T(t)(r(t)-\theta(t))dt)^2] \\ \leq & 3M_a^2C^2r_0^2 + 12M^2C^2 + \frac{6C^2\bar{M}^2}{a^2}. \end{aligned}$$

□

Proof of Theorem 5.5

The idea is to study the expression $\hat{\theta}_{i,T} - \theta_{i,T}$ and derive its limiting distribution. By (5.7),

$$\begin{aligned} & \hat{a}_{T,K_T}(\hat{\Phi}_{T,K_T}(\theta) - \Phi_{T,K_T}(\theta)) - (a - \hat{a}_{T,K_T})\Phi_{T,K_T}(\theta - r) \\ = & A_{T,K_T}^{-1}\int_0^T \frac{\vec{p}_{T,K_T}(t)}{\sigma(t)}dW(t) = \frac{2}{T}\int_0^T \frac{\vec{p}_{T,K_T}(t)}{\sigma(t)}dW(t). \end{aligned}$$

Therefore, for each fixed $i <= K_T$, we have

$$\begin{aligned} & \hat{a}_{T,K_T}(\hat{\theta}_{i,T} - \theta_{i,T}) - (a - \hat{a}_{T,K_T})(\theta_{i,T} - \frac{2}{T}\int_0^T \frac{p_{i,T}(t)r(t)}{\sigma^2(t)dt}dt) \\ = & \frac{2}{T}\int_0^T q_{i,T}(t)dW(t). \end{aligned} \tag{5.34}$$

We emphasize that although $\theta_{i,T} = \frac{2}{T}\int_0^T \frac{q_{i,T}(t)\theta(t)}{\sigma(t)}dt$ is independent of K_T, by (5.5), the term $\hat{\theta}_{i,T} = \frac{2}{T}\int_0^T \frac{q_{i,T}(t)r(t)}{\sigma(t)}dt + \frac{2}{T\hat{a}_{T,K_T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}dr(t)$ depends on K_T. Multiplying both sides of (5.34) by $\sqrt{T}$, we obtain

$$\begin{aligned} & \sqrt{T}\hat{a}_{T,K_T}(\hat{\theta}_{i,T} - \theta_{i,T}) - (a - \hat{a}_{T,K_T})\frac{2}{\sqrt{T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t) - r(t))dt \\ = & \frac{2}{\sqrt{T}}\int_0^T q_{i,T}(t)dW(t), \end{aligned} \tag{5.35}$$

or equivalently

$$\begin{aligned} & \sqrt{T}\hat{a}_{T,K_T}(\hat{\theta}_{i,T} - \theta_{i,T}) \\ = & (a - \hat{a}_{T,K_T})\frac{2}{\sqrt{T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t) - r(t))dt + \frac{2}{\sqrt{T}}\int_0^T q_{i,T}(t)dW(t). \end{aligned}$$

We will show that the first term on the right hand-side of the above equation converges in probability to 0, and the second term converges in distribution to $N(0, 2)$. Then the statement (5.18) will follow by a simple application of Slutsky's theorem. For the second term, we know that $\sqrt{\frac{2}{T}}\int_0^T q_{i,T}(t)dW(t) \sim N(0, 1)$, which follows from $\int_0^T q_{i,T}^2(t)dt = \frac{T}{2}$. Therefore,

$$\frac{2}{\sqrt{T}}\int_0^T q_{i,T}(t)dW(t) \sim N(0, 2).$$

For the first term, under condition $\mathcal{A}8$, it follows from Lemma 5.5 that

$$\sup_{T>0} E[(\frac{2}{\sqrt{T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t) - r(t))dt)^2] < \infty,$$

which implies that $\frac{2}{\sqrt{T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t) - r(t))dt$ is uniformly bounded in probability for $T > 0$. Therefore, under condition $\mathcal{A}9$, elementary probability rules and Slutsky's

theorem imply that

$$(a-\hat{a}_{T,K_T})\frac{2}{\sqrt{T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt \xrightarrow{P} 0. \tag{5.36}$$

We now proceed to prove the statement (5.19). By Itô's isometry, $\int_0^T \frac{q_{i,T}(t)}{\sqrt{T/2}}dW(t), i=0,1,\cdots$ are i.i.d $\sim N(0,1)$. Therefore,

$$\sum_{i\in F_J}(\frac{q_{i,T}(t)}{\sqrt{T/2}}dW(t))^2 \sim \chi^2(J). \tag{5.37}$$

From (5.35),

$$\begin{aligned}
&\sum_{i\in F_J}(\sqrt{2/T}\int_0^T q_{i,T}(t)dW(t))^2 \\
=\ &\sum_{i\in F_J}(\sqrt{T/2}\hat{a}_{T,K_T}(\hat{\theta}_{i,T}-\theta_{i,T})-(a-\hat{a}_{T,K_T})\sqrt{2/T}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt)^2 \\
=\ &\sum_{i\in F_J}\frac{T}{2}\hat{a}^2_{T,K_T}(\hat{\theta}_{i,T}-\theta_{i,T})^2+\sum_{i\in F_J}((a-\hat{a}_{T,K_T})\sqrt{2/T}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt)^2 \\
&-2\sum_{i\in F_J}(\sqrt{T/2}\hat{a}_{T,K_T}(\hat{\theta}_{i,T}-\theta_{i,T})(a-\hat{a}_{T,K_T})\sqrt{2/T}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt.
\end{aligned}$$

By rearranging the above equation, we get

$$\begin{aligned}
&\sum_{i\in F_J}\frac{T}{2}\hat{a}^2_{T,K_T}(\hat{\theta}_{i,T}-\theta_{i,T})^2 \\
=\ &\sum_{i\in F_J}(\sqrt{2/T}\int_0^T q_{i,T}(t)dW(t))^2-\sum_{i\in F_J}((a-\hat{a}_{T,K_T})\sqrt{2/T}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt)^2 \\
&+2\sum_{i\in F_J}(\sqrt{T/2}\hat{a}_{T,K_T}(\hat{\theta}_{i,T}-\theta_{i,T})(a-\hat{a}_{T,K_T})\sqrt{2/T}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt.
\end{aligned}$$

From (5.37), the first term on the right hand-side of the above equation has distribution $\chi^2(J)$. In light of (5.36), the continuous mapping theorem and Slutsky's theorem imply that the second term converges to zero in probability. Using (5.18), (5.36) and Slutsky's theorem, the third term converges in probability

to 0. Thus, the statement (5.19) follows from Slutsky's theorem. □

Proof of Theorem 5.6

The idea in the proof is similar to that for Theorem 5.5. By (5.8),

$$\begin{aligned}
& \hat{a}_{T,K_T}(\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)) \\
= \; & (a - \hat{a}_{T,K_T}) M_{T,K_T}(\theta - r)(t) + \vec{p'}_{T,K_T}(t) A^{-1}_{T,K_T} \int_0^T \frac{\vec{p}_{T,K_T}(t)}{\sigma(t)} dW(t) \\
= \; & (a - \hat{a}_{T,K_T}) \vec{p'}_{T,K_T} A^{-1}_{T,K_T} \int_0^T \frac{\vec{p}_{T,K_T}(t)(\theta - r)(t)}{\sigma^2(t)} dt + \vec{p'}_{T,K_T}(t) A^{-1}_{T,K_T} \int_0^T \frac{\vec{p}_{T,K_T}(t)}{\sigma(t)} dW(t),
\end{aligned}$$

where the last equality follows from the definition of the operator M_{T,K_T}. Using the above result and the facts that $\vec{p}_{T,K_T}(t) = \sigma(t)\vec{q}_{T,K_T}(t)$ and $A_{T,K_T} = \frac{T}{2} I_{(K_T+1)\times(K_T+1)}$, we have

$$\begin{aligned}
& \hat{a}_{T,K_T}(\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)) \\
= \; & (a - \hat{a}_{T,K_T}) \sum_{i=0}^{K_T} \sigma(t) q_{i,T}(t) \frac{2}{T} \int_0^T \frac{q_{i,T}(t)}{\sigma(t)} (\theta(t) - r(t)) dt \\
& + \sigma(t) \frac{2}{T} \sum_{i=0}^{K_T} q_{i,T}(t) \int_0^T q_{i,T}(t) dW(t).
\end{aligned}$$

Multiplying both sides of the above equation by $\sqrt{\frac{T}{\sum_{i=0}^{K_T} q^2_{i,T}(t)}}$, we get

$$\begin{aligned}
& \sqrt{\frac{T}{\sum_{i=0}^{K_T} q^2_{i,T}(t)}} \hat{a}_{T,K_T}(\hat{M}_{T,K_T}(\theta)(t) - M_{T,K_T}(\theta)(t)) \\
= \; & (a - \hat{a}_{T,K_T}) \sqrt{\frac{T}{\sum_{i=0}^{K_T} q^2_{i,T}(t)}} \sum_{i=0}^{K_T} \sigma(t) q_{i,T}(t) \frac{2}{T} \int_0^T \frac{q_{i,T}(t)}{\sigma(t)} (\theta(t) - r(t)) dt \\
& + \sqrt{\frac{T}{\sum_{i=0}^{K_T} q^2_{i,T}(t)}} \sigma(t) \frac{2}{T} \sum_{i=0}^{K_T} q_{i,T}(t) \int_0^T q_{i,T}(t) dW(t).
\end{aligned}$$

We prove that the first term on the right hand-side of the above equation converges in probability to 0 and the second term converges in distribution to $N(0, 2\sigma^2(t))$.

Then the statement of our theorem follows from Slutsky's theorem. For the second term, we have

$$\begin{aligned} & \sqrt{\frac{T}{\sum_{i=0}^{K_T} q_{i,T}^2(t)}}\sigma(t)\frac{2}{T}\sum_{i=0}^{K_T} q_{i,T}(t)\int_0^T q_{i,T}(t)dW(t) \\ = \quad & \sqrt{2}\sigma(t)\frac{\sum_{i=0}^{K_T} q_{i,T}(t)\int_0^T \frac{q_{i,T}(t)}{\sqrt{T/2}}dW(t)}{\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)}}. \end{aligned}$$

By the Itô isometry, the variables $\int_0^T \frac{q_{i,T}(t)}{\sqrt{T/2}}dW(t), i = 0, 1, \cdots$ are i.i.d $\sim N(0,1)$. Therefore,

$$\frac{\sum_{i=0}^{K_T} q_{i,T}(t)\int_0^T \frac{q_{i,T}(t)}{\sqrt{T/2}}dW(t)}{\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)}} \sim N(0,1),$$

which gives

$$\sqrt{2}\sigma(t)\frac{\sum_{i=0}^{K_T} q_{i,T}(t)\int_0^T \frac{q_{i,T}(t)}{\sqrt{T/2}}dW(t)}{\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)}} \sim N(0, 2\sigma^2(t)).$$

For the first term,

$$\begin{aligned} & (a-\hat{a}_{T,K_T})\sqrt{\frac{T}{\sum_{i=0}^{K_T} q_{i,T}^2(t)}}\sum_{i=0}^{K_T}\sigma(t)q_{i,T}(t)\frac{2}{T}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt \\ = \quad & \sigma(t)(a-\hat{a}_{T,K_T})\sqrt{\sum_{i=0}^{K_T} C_i^2}\frac{\sum_{i=0}^{K_T} q_{i,T}(t)\frac{2}{\sqrt{T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt}{\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)\sum_{i=0}^{K_T} C_i^2}}. \end{aligned}$$

We shall prove that $\frac{\sum_{i=0}^{K_T} q_{i,T}(t)\frac{2}{\sqrt{T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt}{\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)\sum_{i=0}^{K_T} C_i^2}}$ is un6iformly bounded in probability by showing that its second moment is finite. Then, under condition $\mathcal{A}10$, it follows from Slutsky's theorem that the term

$$\sigma(t)(a-\hat{a}_{T,K_T})\sqrt{\sum_{i=0}^{K_T} C_i^2}\frac{\sum_{i=0}^{K_T} q_{i,T}(t)\frac{2}{\sqrt{T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt}{\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)\sum_{i=0}^{K_T} C_i^2}}$$

converges to 0 in probability. Now we proceed to prove that

$$\sup_{T>0} E[(\frac{\sum_{i=0}^{K_T} q_{i,T}(t)\frac{2}{\sqrt{T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt}{\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)\sum_{i=0}^{K_T} C_i^2}})^2] < \infty.$$

By the Cauchy-Schwartz inequality,

$$\begin{aligned} & E[(\frac{\sum_{i=0}^{K_T} q_{i,T}(t)\frac{2}{\sqrt{T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt}{\sqrt{\sum_{i=0}^{K_T} q_{i,T}^2(t)\sum_{i=0}^{K_T} C_i^2}})^2] \\ \leq & \frac{\sum_{i=0}^{K_T} E(\frac{2}{\sqrt{T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt)^2}{\sum_{i=0}^{K_T} C_i^2}. \end{aligned}$$

Under condition $\mathcal{A}8$, we have by Lemma 5.5

$$\begin{aligned} & \frac{\sum_{i=0}^{K_T} E(\frac{2}{\sqrt{T}}\int_0^T \frac{q_{i,T}(t)}{\sigma(t)}(\theta(t)-r(t))dt)^2}{\sum_{i=0}^{K_T} C_i^2} \\ \leq & \frac{\sum_{i=0}^{K_T} \frac{2C_i^2}{\bar{\epsilon}^2}(3M_a^2 r_0^2 + 12M^2 + \frac{6\bar{M}^2}{a^2})}{\sum_{i=0}^{K_T} C_i^2} = \frac{6M_a^2 r_0^2 + 24M^2 + \frac{12\bar{M}^2}{a^2}}{\bar{\epsilon}^2}, \end{aligned}$$

which is uniformly bounded for $T > 0$. Note that the inequality follows from our assumption that $\sigma(t) > \bar{\epsilon} > 0$. □

Chapter 6

Summary and Future Research

Stochastic processes are widely applied in engineering, finance, and economics. Given observed data of underlying objects, statistical inference is useful for model building and hence for understanding the behavior of the stochastic evolution of modeling objects. It is reasonable to believe that the parameters governing the stochastic processes are time dependent. In finance, there are typically two sources of information available: historical data under P measure or the objective measure, and market data under Q measure or risk-neutral measure. It is advisable to make use of information from both sources. However, the model parameterizations are usually different under P and Q measure. Actually, the models under P measure are typically time independent and those under Q measure are typically time dependent. Motivated by papers by Fan et. al (2003) and Al-Zoubi (2009), this thesis has focused on statistical inference for time-inhomogeneous SDEs. We have assumed that a continuous realization of the SDE is available.

In Chapter 2, we have proposed a new class of time-dependent SDEs by allowing the parameters to change with the time elapsed in each regime. Due to the latent regime-switching process, we do not have a closed-form expression of the likelihood function of unknown parameters. We discretize the continuous process with an

Euler scheme and use an EM algorithm (Hamilton, 1990) to estimate parameters. Our simulation studies show that the parameters can be efficiently estimated. In principle, this estimation procedure can be applied to any TDRS model. One future research direction is to develop asymptotic properties of the resulting maximum likelihood estimator, such as the asymptotic consistency and normality.

In Chapter 3, we proved that the TDRS Vasicek model is stationary. The proof relies on an explicit decomposition of the model's solution. One may generalize the results to a more general setting, i.e., for general TDRS models. In that case, tools such as those presented in Mao and Yuan (2006) may be necessary.

In Chapter 4, we studied a Brownian motion with time-dependent drift $\theta(t)$, and derived a maximum likelihood estimator for the projection of $\theta(t)$ onto a finite dimensional space $V_{T,K}$. Using a sieve-type method, we have proven that the proposed maximum likelihood estimator is asymptotically consistent as long as the dimension of parameter space does not increase very fast. The objective in this thesis has been to estimate $M_{T,K}(\theta)(t)$, the projection of $\theta(t)$ onto $V_{T,K}$. It is still an open question under what conditions $\hat{M}_{T,K_T}(\theta)(t)$ converges in probability to $\theta(t)$. This is certainly an interesting research problem.

In Chapter 5, we generalized the results in Chapter 4 to a class of mean-reverting SDEs with the time-dependent mean-reversion level function $\theta(t)$. The estimation objective is the same as in Chapter 4, i.e., to estimate $M_{T,K}(\theta)(t)$. Since the mean-reverting speed parameter is unknown, the derivations have been much more involved. We have proven the asymptotic consistency and normality of the proposed maximum likelihood estimator $\hat{M}_{T,K_T}(\theta)(t)$. Moreover, in finite sample case, we have proposed a heuristic approach to determine the dimension of parameter space and construct a confidence interval for $M_{T,K}(\theta)(t)$ for a given length of data T. A future research direction will be to simplify the technical conditions in the theorems so that they are easier to check in practice.

Bibliography

[1] Aït-Sahalia, Y. (1996). Testing continuous-time models of the spot interest rate. *Review of Financial Studies* **9**, 385–426.

[2] Aït-Sahalia, Y. (1999). Transition densities for interest rate and other nonlinear diffusions. *J. Finance* **LIV** , 1361–1395.

[3] Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrika* **70**, 223-262.

[4] Al-Zoubi, H. A. (2009). Short-term spot rate models with nonparametric deterministic drift. *The Quarterly Review of Economics and Finance* **49**, 731–747.

[5] Anderson, W.J. (1991). *Continuous-Time Markov Chains: An Applications-Oriented Approach*. Springer-Verlag.

[6] Andrews, D. (1991b). Asymptotic normality of series estimators for nonparametric and semiparametric regression models *Econometrica* **59**, 307–345.

[7] Ang, A. and Bekaert, G. (2002). Regime switches in interest rates. *Journal of Business and Economic Statistics* **20**, 163–182.

[8] Ashley, R. and Patterson, D.M. (2007). Apparent long memory in time series as an artifact of a time-varying mean: a "local mean" filtering alternative to

the fractionally integrated model. Mimeo, Economics Department, Virginia Tech.

[9] Ashley, R. and Patterson, D. M. (2010). Apparent long memory in time series as an artifact of a time-varying mean: considering alternatives to the fractionally integrated model. *Ma5croeconomic Dynamics* **14**, (Supplement 1), 59–87.

[10] Ball, C. and Torous, W. (1996). Unit roots and the estimation of interest rate dynamics. *Journal of Empirical Finance* **3**, 215–238.

[11] Bansal, R. and Zhou, H. (2002). Term structure of interest rates with regime shifts. *Journal of Finance* **57**, 1997–2043.

[12] Barberá, S., Hammond, P. J. and Seidl, C. (2003). *Handbook of Utility Theory.* Kluwer Academic Publishers, Boston.

[13] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic function of finite state Markov chains. *Annals of Mathematical Statistics* **30**, 1554–1563.

[14] Baxter, M. and King, R. (1999). Measuring business cycles: Approximate band-pass filters for economic time series. *Review of Economics and Statistics* **81**, 575–593.

[15] Beder, J. H., (1987). A sieve estimator for the mean of gaussian process. *Ann. Statist.* **15**, 59–78.

[16] Bickel, P. J., Ritov, Y. and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist* **26 (4)**, 1614–1635.

[17] Bierens, H. J. (1997) Testing the unit root with drift hypothesis against nonlinear trend stationarity, with an application to the U.S. price level and interest rate. *Journal of Econometrics* **81**, 29–64.

[18] Bierens, H. J. and Carvalho, J. R. (2007). Semi-nonparametric competing risks analysis of recidivism. *Journal of Applied Econometrics* **22**, 971–993.

[19] Bierens, H. J. (2008). Semi-nonparametric interval-censored mixed proportional hazard models: identification and consistency results. *Econometric Theory* **24**, 749–794.

[20] Bierens, H. J. (2011). Consistency and asymptotic normality of sieve estimators under weak and verifiable conditions. http://econ.la.psu.edu/hbierens/snpmodels.pdf.

[21] Bishwal, J. P. N. (2008). *Parameter Estimation in Stochastic Differential Equations*. Springer-Verlag.

[22] Black, F., Derman, E. and Toy, W. (1990). A one-factor model of interest rates and its application to treasury bond options. *Financial Analysts Journal*, January–February, 33–39.

[23] Black, F. and Karasinski, P. (1991). Bond and option pricing when short rates are lognormal. *Financial Analysts Journal*, July–August, 52–59.

[24] Blundell, R., Chen, X., and Kristensen, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* **75 (6)**, 1613–1669.

[25] Bougerol, P., and Picard, N . (1992). Stationarity of GARCH processes and of some non-negative time series. *Journal of Econometrics* **52 (1-2)**, 115–127.

[26] Brandt, A. (1986). The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Adv. Appl. Probab.* **18**, 211–220.

[27] Brandt, M. W. and Santa-Clara, P. (2002). Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets. *Journal of Financial Economics* **63**, 161–210.

[28] Brennan, M. J. and Schwartz, E. S. (1979). A continuous time approach to the pricing of bonds. *Journal of Banking and Finance* **3(2)**, 133–155.

[29] Brigo, D. and Merurio, F. (2001). *Interest Rate Models: Theory and Practice*. Springer.

[30] Cai, Z. and Hong, Y. (2009). Some recent developments in nonparametric finance. *Advances in Econometrics* **25**, 379–432.

[31] Cairns, A. J. G. (2004). *Interest Rate Models: An Introduction*. Princeton University Press.

[32] Campbell, J. Y. (1986). A defense of traditional hypotheses about the term structure of interest rates. *Journal of Finance* **41**, 183–193.

[33] Canuto, C., Hussaini, M. Y., Quarteroni, A., and Zang, T. A. (2006). *Spectral methods: fundamentals in single domains*. Springer.

[34] Chan, K. C., Karolyi, G. A., Longstaff, F. A., and Sanders, A. B. (1992). An Empirical Comparison of Alternative Models of the Short-Term Interest Rate. *Journal of Finance* **47**, 1209–1227.

[35] Chapman, D. A. and Pearson, N. D. (2000). Is the short rate drift actually nonlinear ? *Journal of Finance* **55**, 355–388.

[36] Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics* (J. J. Heckman and E. E. Leamer, eds.) 6B 5549–5632. North Holland, Amsterdam.

[37] Chen, X., Fan, Y., and Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association* **101**, 1228–1240.

[38] Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica* **66**, 289–314.

[39] Chen, Z. L. and Forsyth, P. A. (2010). Implications of a regime-switching model on natural gas storage valuation and optimal operation. *Quantitative Finance* **10 (2)**, 159–176.

[40] Chernov, M. (2001). Implied volatilities as forecasts of future volatility, time-varying risk premia, and returns variability, AFA 2002 Atlanta Meetings.

[41] Chernozhukov, V., Imbens, G. and Newey, W. (2007). Instrumental variable identification and estimation of nonseparable models via quantile conditions. *Journal of Econometrics* **139**, 4–14.

[42] Choi, S. (2002). Maximum likelihood estimation of continuous-time diffusion process model with nonlinear drift and constant elasticity of volatility for short-term interest rate. Working Paper.

[43] Choi, S. (2009). Regime-switching univariate diffusion models of the short-term interest rate. *Studies in Nonlinear Dynamics and Econometrics* **13**, No. 1, Article 4.

[44] Cogley, T. and Sbordone, A. M. (2008). Trend Inflation, Indexation, and Inflation Persistence in the New Keynesian Phillips Curve. *American Economic Review* **98:5**, 2101–2126.

[45] Cox, J. C., Ingersoll, J. E. and Ross, S. A. (1985). A Theory of the Term Structure of Interest Rates. *Econometrica* **53**, 385–408.

[46] Darolles, S. and Gourieroux, C. (2001). Truncated dynamics and estimation of diffusion equations. *Journal of Econometrics* **102 (1)**, 1–22.

[47] Davis, M. H. A. (1993). *Markov Models and Optimization*. Chapman & Hall.

[48] Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Application*, Cambridge University Press.

[49] Dehling, H., Franke, B. and Kott, T. (2010) Drift estimation for a periodic mean reversion process. *Stat Inference Stoch Process* **13**, 175–192.

[50] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm *Journal of the Royal Statistical Society B* **39**, 1–38.

[51] Douc, R., Moulines, E. and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *The Annals of Statistics* **32**, 2254-2304.

[52] Driffill, J., Kenc, T. and Sola, M. (2002). An empirical examination of term structure models with regime shifts. Working paper.

[53] Durham, G. B. (2003). Likelihood-based specification analysis of continuous-time models of the short-term interest rate. *Journal of Financial Economics* **70**, 463–487.

[54] Durham, G. B. and Gallant, A. R. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics* **20 (3)**, 297–316.

[55] Efron, B., (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7 (1)**, 1–26.

[56] Engle, R., Granger, C., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* **81**, 310–320.

[57] Engle, R. and Rangel, G. (2004). The spline GARCH model for unconditional volatility and its global macroeconomic causes. Working paper. New York University.

[58] Fan, J. and Gijbels I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society. Series B(Methodological)* **57**, 371-394.

[59] Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645-660.

[60] Fan, J. and Zhang, C. (2001). A re-examination of Stanton's diffusion estimations with applications to financial model validation. *Journal of the American Statistical Association* **97** , 118-134.

[61] Fan, J., Jiang, J., Zhang, C. and Zhou, Z. (2003). Time-dependent diffusion models for term structure dynamics. *Statistica Sinica* **13**, 965-992.

[62] Fan, J. (2005). A selective overview of nonparametric methods in financial econometrics. *Statistical Science* **20**, 317-337.

[63] Fan, J., Fan, Y. and Jiang, J. (2007). Dynamic integration of time- and state-domain methods for volatility estimation. *Journal of the American Statistical Association* **102**, 618-631.

[64] Fong, H. G. and Vasicek, O. A. (1991). Fixed income volatility management. *The Journal of Portfolio Management Summer* **2**, 41–46.

[65] Francq, C. and Roussignol, M. (1998). Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum-likelihood estimator. *Statistics: A Journal of Theoretical and Applied Statistics* **32**, 151-173.

[66] Francq, C. and Zakoïan, J. (2001). Stationarity of multivariate Markov-switching ARMA models. *Journal of Econometrics* **102**, 339–364.

[67] Gallant, A. R. and Souza, G. (1991). On the asymptotic normality of Fourier flexible form estimates. *Journal of Econometrics* **50**, 329–353.

[68] Gallant, A. R. and Tauchen, G. (1996). Which moments to match? *Econometric Theory* **12**, 657–681.

[69] Gallant, A. R. and Tauchen, G. (2004). EMM: A program for efficient method of moments estimation, Version 2.0 User Guide. Working paper. Duke University.

[70] Garcia, R. and Perron, P. (1996). An analysis of the real interest rate under regime shifts *Review of Economics and Statistics* **78**, 111–125.

[71] Geman, S. and Hwang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics* **10**, 401–414.

[72] Genon-Catalot, V., Laredo, C. and Picard, D. (1992). Non-parametric estimation of the diffusion coefficient by wavelets methods. *Scandinavian Journal of Statistics* **19**, 317–335.

[73] Gland, F. L. and Mevel, L. (2000). Exponential forgetting and geometric ergodicity in Hidden Markov models. *Math. Control Signals Systems* **13**, 63–69.

[74] Gospodinov, N., Gavala, A. and Jiang, D. (2006). Forecasting volatility. *Journal of Forecasting* **25**, 381–400.

[75] Gray, S. (1996), Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics* **42**, 27–62.

[76] Grenander, U. (1981). *Abstract Inference*. Wiley, New York.

[77] Hamilton, J. D. (1988). Rational-expectations econometric-analysis of changes in regime - an investigation of the term structure of interest-rates. *Journal of Economic Dynamics & Control* **12 (2-3)**, 385–423.

[78] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357–384.

[79] Hamilton, J. D. (1990). Analysis of time series subject to change in regime. *Journal of Econometrics* **45**, 39–70.

[80] Hamilton, J. D. (1994). *Time Series Anlasis*. Princeton University Press.

[81] Hamming, R. W. (1973). *Numerical Methods for Scientists and Engineers*, New York: Dover Publications.

[82] Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50 (3)**, 1029–1054.

[83] Hardy, M. (2001). A regime-switching model of long term-stock returns. *North American Actuarial Journal* **5.2**, 41–53

[84] Hart, J. (1997).*Nonparametric Smoothing and Lack-of-Fit Tests*. Springer-Verlag, New York.

[85] Heath, D., Jarrow, R. and Morton, A. (1992). Bond pricing and the term structure of interest rates. *Econometrica* **60**, 77–106.

[86] Heston, S. L. (1993). A Closed-form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *Review of Financial Studies* **6 (2)**, 327–343

[87] Hyndman, R. J., Koehler, A. B., Ord, J. K and Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer-Verlag, Berlin Heidelberg.

[88] Ho, T. and Lee, S. (1986). Term Structure Movements and Pricing Interest Rate Contingent Claims. *Journal of Finance* **41**, 1011–1029.

[89] Hofmann, B., Kramer, R. and Richter, M. (2009). Some aspects of parameter identification in a mean reverting financial asset model with time-dependent volatility. *International Journal of Computer Mathematics* **86 (6)**, 992–1008.

[90] Hong, Y. and White, H. (1995). Consistent specification testing via nonparametric series regression. *Econometrica* **63**, 1133–1159.

[91] Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics* **31**, 1600–1635.

[92] Hull, J. and White, A. (1987). The Pricing of Options on Assets with Stochastic Volatilities. *Journal of Finance* **42 (2)**, 281–300.

[93] Hull, J. and White, A. (1994a). Numerical procedures for implementing term structure models I: Single-Factor Models. *Journal of Derivatives* **2,1**, (Fall 1994a) 7–16.

[94] Imbens, G., Newey, W. K. and Ridder, G. (2005). Mean-squared-error calculations for average treatment effects. Manuscript. UC Berkeley

[95] James, J. and Webber, N. (2000). *Interest Rate Modelling*. John Wiley & Sons, Ltd.

[96] Jiang, G. J. and Knight, J. L. (1997). Nonparametric Approach to the Estimation of Diffusion Processes - With an Application to a Short-Term Interest Rate Model. *Econometric Theory* **13**, 615-645.

[97] Karlsen, H. A. (1990). Doubly stochastic vector AR(1) processes. Technical Report, Dept. of Mathematics, Univ. of Bergen, Norway.

[98] Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer-Verlag.

[99] Kim, C. (1993). Dynamic linear models with Markov-switching. *Journal of Econometrics* **60**, 1–22.

[100] Kloeden, P. E. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Springer, New York.

[101] Krishnamurthy, V. and Rydén, T. (1998). Consistent estimation of linear and nonlinear autoregressive models with Markov regime. *J. Time Ser. Anal.* **19** 291–307.

[102] Kutoyants, Yu. A. (1984). Parameter estimation for diffusion type processes of observation. *Math. Operationsforsch. u. Statist., ser. statist.* **4**, 541–551.

[103] Lampard, D. G. and Redman, S. J. (1963). Statistical properties of the integral of a binary random process. *Circuits Theory, IEEE Transactions* **10**, 413–427.

[104] Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.* **40**, 127–143.

[105] Leśkow, J. and Różański, R. (1989). Maximum likelihood estiamtor of the drift function for a diffusion process. *Statistics and Decisions* **7**, 243–262.

[106] Li, Q., Hsiao, C. and Zinn, J. (2003). Consistent specification tests for semiparametric/nonparametric models based on series estimation methods. *Journal of Econometrics* **112**, 295–325.

[107] Lipster, R. S. and Shiryaev, A. N. (2001). *Statistics of Random Processes: General theory*. Springer-Verlag.

[108] Lipster, R. S. and Shiryaev, A. N. (2010). *Statistics of Random Processes: II. Applications*. Springer-Verlag

[109] Longstaff, F. and Schwartz, E. S. (1992). Interest rate volatility and the term structure: A two-factor general equilibrium model. *Journal of Finance* **47**, 1259–1282.

[110] Mackeague, I. W. (1986). Estimation for a Semimartingale Regression-model Using the Method of Sieves. *Annals of Statistics* **14 (2)**, 579–589.

[111] Mackeague, I. W. (1988). The method of sieves. *Enclopedia of Statistical Sciences* **8**, 458–461.

[112] Mao, X. and Yuan, C. (2006). *Stochastic Differential Equations with Markovian Switching*. Imperial College Press.

[113] McFadden, J. A. (1959). The probability density of the output of an RC filter when the input is a binary random process. *IRE Trans. on Information Theory* **5**, 174–178.

[114] Mclachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Willey Series.

[115] McLeish, D. L. and Kolkiewicz, A. (1997). Fitting diffusion models in finance: Selected Proceedings of the Conference on Estimating Functions. *Ins. Math. Stat. Lect. Notes*, 309–332.

[116] McLeish, D. L. and Small, C. G. (1988). *The Theory and Applications of Statistical Inference Functions.* Lecture Notes in Statistics 44, Springer-Verlag, New York.

[117] Misscia, O. D. (2004). Nonparametric estmation of diffusion process: a closer look. Working Paper.

[118] Naik, V. and Lee, M. H. (1997). Yield curve dynamics with discrete shifts in economic regimes: theory and estimation. Working Paper, Faculty of Commerce, University of British Columbia.

[119] Newey, W. K. (1994b). Series estimation of regression functionals. *Econometric Theory* **10**, 1–28.

[120] Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* **79**, 147–168.

[121] Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71**, 1565–1578.

[122] Newey, W. and West, K. (1987) Hypothesis testing with efficient method of moments estimation. *International Economic Review* **28**, 777–787.

[123] Nguyen, H. T. and Pham, D. T. (1982). Identication of nonstationary diffusion model by the method of sieves. *SIAM J. Control Optimiz.* **20**, 603–611.

[124] Pedersen, A. R. (1995a). A new approach to maximum likelihood estimation for stochastic differential equations. *Scandinavian Journal of statistics* **22**, 55–71.

[125] Pedersen, A. R. (1995b). Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli* **1**, 257–279.

[126] Phillips, P. C. and Yu, J. (2005). Jackknifing bond option prices. *Review of Financial Studies* **18**, 707–742.

[127] Prakasa Rao, B. L. S. (1999). *Statistical Inference for Diffusion Type Processes*. Arnold.

[128] Prakasa Rao, B. L. S. (2004). Identification for linear stochastic systems driven by fractional brownian motion. *Stochastic Analysis and Applications* **22 (6)**, 1487–1509.

[129] Redekop, J. and Wirjanto, T. S. (2010), Exploring a Two-State Markov-Chain Model for Option-Pricing, Unpublished Manuscript, University of Waterloo, April.

[130] Redekop, J. and Wirjanto, T. S. (2010), On the Moment Generating Functions of a Two-State Markov-Chain Option-Pricing Model, Unpublished Manuscript, University of Waterloo, June.

[131] Santa-Clara, P. (1995). Simulated likelihood estimation of diffusions with an application to the short-term interest rate. Working paper, Anderson Graduate School of Management, UCLA.

[132] Satish, L. and Gururaj, B. (April 2003). Use of hidden Markov models for partial discharge pattern classification. *IEEE Transactions on Dielectrics and Electrical Insulation*.

[133] Song, K. (2005). Testing semiparametric conditional moment restrictions using conditional martingale transforms. Manuscript. Yale University, Department of Economics.

[134] Sørensen, H. (2003). Simulated likelihood approximations for stochastic volatility models. *the Scandinavian Journal of Statistics* **30**, 257–276.

[135] Stanton, R. (1997). A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk. *The Journal of Finance* **5**, 1973-2002.

[136] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer-Verlag.

[137] Stinchcombe, M. and White, H. (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory* **14**, 295–325.

[138] Stone, C. J. (1990). Large-sample inference for log-spline models. *The Annals of Statistics* **18**, 717–741.

[139] Stone, C. J. and Huang, J. Z. (2003). Statistical modeling of diffusion processes with free knot splines. *Journal of Statistical Planning and Inference* **116**, 451–474.

[140] Strawderman, R. L. and Tsiatis, A. A. (1996). On the asymptotic properties of a flexible hazard estimator. *The Annals of Statistics* **24**, 41–63.

[141] Takamizawa, H. (2008). Is Nonlinear Drift Implied by the Short End of the Term Structure? *Review of Financial Studies* **21(1)**, 311–346.

[142] Tang, C. Y. and Chen, S. X. (2009). Parameter estimation and bias correction for diffusion processes. *Journal of Econometrics* **149**, 65–81.

[143] Thorisson, H. (2000). *Coupling, Stationarity and Regeneration*. Springer-Verlag.

[144] Tuan, P. (1981). Nonparametric estimation of the drift coefficient in the diffusion equation. *Math. Operationsforsch. Statist. Ser. Statistics* **12 (1)**, 61–73.

[145] Van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *The Annals of Statistics* **21**, 14–44.

[146] VASICEK, O. A. (1977). An Equilibrium Characterization of the Term Structure. *Journal of Financial Economics* **5**, 177–88.

[147] White, H. and Wooldridge, J. (1991). Some results on sieve estimation with dependent observations. In: Barnett, W.A., Powell, J., and Tauchen, G. (Eds.), Non-parametric and Semi-parametric Methods in Econometrics and Statistics. Cambridge University Press, Cambridge, pp. 459–493.

[148] Wirjanto, T. S. (2006), On the Estimation of Two-state Markov-Chain Process of Short-Term Interest Rate, Unpublished Manuscript, University of Waterloo, January.

[149] Wirjanto, T. S. (2010), A Specification Test for Diffusion Processes, Unpublished Manuscript, University of Waterloo, April.

[150] Wirjanto, T. S. (2010), A Selected Review of Statistical Tests for Diffusion Processes, Unpublished Manuscript, University of Waterloo, December.

[151] Wonham, W. M. and Fuller, A. T. (1958). Probability densities of the smoothed random telegraph signal. *J. Electronics and Control* **4**, 567–576.

[152] Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates for sieve MLE. *The Annals of Statistics* **23**, 339–362.

[153] Yan, H. (2001). Dynamic models of the term structure. *Financial Analysts Journal* **57**, 60–76.

[154] Yu, J. and Phillips, P. C. (2001). A Gaussian approach for estimating continuous time models of short term interest rates *The Econometrics Journal* **4**, 211–225.

[155] Yu, J. (2011). Bias in the estimation of the mean reversion parameter in continuous time models *Journal of Econometrics*, forthcoming.

[156] Zhao, Z. (2008). Parametric and nonparametric models and methods in financial econometrics. *Statistical Surveys* **2**, 1–42.

[157] Zhou, S., Shen, X. and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics* **26**, 1760–1782.